DOCUMENT RESUME

ED 081 263 FL 004 376

TITLE	Speech Research: A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for Its Investigation, and Practical Applications, 1 January - 31 March 1973.
INSTITUTION	Haskins Labs., New Haven, Conn.
SPONS AGENCY	National Inst. of Child Health and Human Development (NIH), Bethesda, Md.; National Inst. of Dental Research (NIH), Bethesda, Md.; Office of Naval Research, Washington, D.C. Information Systems Research.
REPORT NO	SR-33-73
PUB DATE	Mar 73
NOTE	269p.
EDRS PRICE	MF-$0.65 HC-$9.87
DESCRIPTORS	Articulation (Speech); Auditory Discrimination; Consonants; Distinctive Features; Language Development; Language Patterns; *Language Research; Linguistics; Memory; Morphology (Languages); *Phonology; *Physiology; Reading Research; *Research Tools; *Speech; Syllables; Vowels

ABSTRACT

This document contains 21 reports on speech research relating to the following areas: phonology, speech development, speech perception, phonetics, short-term memory of tactile stimuli, reading, linguistic and paralinguistic interchange, computer processing of EMG (electromyographic) signals, pitch determination by adaptive autocorrelation method, physiology and speech, and the evolution of language. A list of related publications and reports is also provided. (DD)

FILMED FROM BEST AVAILABLE COPY

ED 081263

SR-33 (1973)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SPEECH RESEARCH

A Report on the Status and Progress of Studies on the Nature of Speech, Instrumentation for its Investigation, and Practical Applications

1 January - 31 March 1973

Haskins Laboratories
270 Crown Street
New Haven, Conn. 06510

Distribution of this document is unlimited.

(This document contains no information not freely available to the general public. Haskins Laboratories distributes it primarily for library use. Copies are available from the National Technical Information Service or the ERIC Document Reproduction Service. See the Appendix for order numbers of previous Status Reports.)

FL 004 376

ACKNOWLEDGMENTS

The research reported here was made possible in part by support from the following sources:

Information Systems Branch, Office of Naval Research
Contract N00014-67-A-0129-0001-0002

National Institute of Dental Research
Grant DE-01774

National Institute of Child Health and Human Development
Grant HD-01994

Research and Development Division of the Prosthetic and Sensory Aids Service, Veterans Administration
Contract V101(134)P-71

National Science Foundation
Grant GS-28354

National Institutes of Health
General Research Support Grant RR-5596

National Institute of Child Health and Human Development
Contract NIH-71-2420

The Seeing Eye, Inc.
Equipment Grant

CONTENTS

I. Manuscripts and Extended Reports

Are You Asking Me, Telling Me, or Talking to Yourself? -- Kerstin Hadding and Michael Studdert-Kennedy 1

Discrimination of Intensity Differences on Formant Transitions In and Out of Syllable Context -- M. F. Dorman 13

Phonological Fusion in Synthetic and Natural Speech -- James E. Cutting . . 19

A Speech Perception Paradox?: The Right-Ear Advantage and the Lag Effect -- Robert A. Weeks 29

Perception of Speech and Nonspeech, with and without Transitions -- James E. Cutting 37

Dichotic Release from Masking for Speech -- Timothy C. Rand 47

Speech Misperception: Inferences About a Cue for Cluster Perception from a Phonological Fusion Task -- James E. Cutting 57

Cross-Language Study of the Perception of the F3 Cue for [r] versus [l] in Speech- and Nonspeech-Like Patterns -- Kuniko Miyawaki, A. M. Liberman, O. Fujimura, Winifred Strange, and J. J. Jenkins 67

Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results) -- P. W. Nye and J. H. Gaitenby . . . 77

Forward and Backward Masking of Brief Vowels -- M. Dorman, D. Kewley-Port, S. Brady-Wood, and M. T. Turvey 93

Effects of Proactive Interference and Rehearsal on the Primary and Secondary Components of Short-Term Retention -- M. T. Turvey and Robert A. Weeks 101

On the Short-Term Retention of Serial, Tactile Stimuli -- Edie V. Sullivan and M. T. Turvey 123

Phonetic Activity in Reading: An Experiment with Kanji -- Donna Erickson, Ignatius G. Mattingly, and Michael T. Turvey 137

Segmentation of the Spoken Word and Reading Acquisition -- Isabelle Y. Liberman 157

Linguistic and Paralinguistic Interchange -- Philip Lieberman 167

Computer Processing of EMG Signals at Haskins Laboratories -- Diane Kewley-Port 173

Pitch Determination by Adaptive Autocorrelation Method -- Georgije Lukatela 185

An Electromyographic Study of the American English Liquids -- David R. Leidner 195

The Role of the Extrinsic and Intrinsic Tongue Muscles in Differentiating the English Tense-Lax Vowel Pairs -- Lawrence J. Raphael and Fredericka Bell-Berti 203

Effect of Speaking Rate on Labial Consonant-Vowel Articulation -- T. Gay, T. Ushijima, H. Hirose, and F. S. Cooper 221

On the Evolution of Language: A Unified View -- Philip Lieberman 229

II. Publications and Reports 271

III. Appendix 275

DDC and ERIC numbers (SR-21/22 - SR-31/32)

I. MANUSCRIPTS AND EXTENDED REPORTS

Are You Asking Me, Telling Me, or Talking to Yourself?

Kerstin Hadding[+] and Michael Studdert-Kennedy[++]

The present study extends earlier work (Studdert-Kennedy and Hadding, 1972) by asking listeners to separate various fundamental frequency contours into three categories rather than two. Its purpose was to establish the perceptual validity and linguistic function of the third category. The fundamental frequency of a 700 msec vocoded utterance, "November" [no'vɛmbə˄], was systematically varied to produce 72 contours differing in f_0 at the stress and over the terminal glide. The contours were recorded (1) carried on the speech wave and (2) as frequency modulated sine waves. Twenty-two Swedish subjects classified (1) both speech and sine wave contours as terminally rising, falling, or level (psychophysical judgments), and (2) speech contours as those of a speaker addressing a question to a listener, making a statement to a listener, or talking to himself (linguistic judgments). The results show that listeners can, with some reliability, separate terminally level from terminally rising or falling contours. If, further, the level terminal glide is combined with an even, low to moderate pitch over earlier sections of the contour, listeners tend to judge the utterance as that of a speaker talking to himself. A new prosodic feature [± Listener] implemented by variations in fundamental frequency is proposed.

In a study of Swedish intonation, Hadding-Koch (1961) distinguished among three functional categories of utterance and their correlated fundamental frequency (f_0) contours. The first category ("question") occurred when a speaker wanted an answer from a listener; it was characterized by a relatively high f_0 at the stress peak and a rising terminal glide. The second category ("statement") occurred when a speaker wanted a listener to believe or agree with him; it was characterized by a lower f_0 at the stress peak and a falling terminal glide. Later perceptual studies of synthetic speech, in which the f_0 contour of an utterance was systematically varied, have largely supported these descriptive analyses (Hadding-Koch and Studdert-Kennedy, 1964, 1965a, 1965b; Studdert-Kennedy and Hadding, 1972, in press). Listeners tended to classify contours with an apparent terminal rise and/or high f_0 at the stress as questions, contours with an apparent terminal fall and/or low f_0 at the stress as statements (cf. Uldall, 1962).

[+]Lund University, Sweden.

[++]Haskins Laboratories, New Haven, Conn.; also Queens College and Graduate Center, University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

The third category of utterance, described by Hadding-Koch, had a level terminal glide ("terminal sustain"). With a relatively even and moderately high overall f_o, this type of contour occurred when the speaker was musing or talking to himself. With various other f_o patterns in earlier sections of the contour, level terminal glides also occurred in exclamations and in some other types of utterance expressing a somewhat emotional reaction. The latter are not considered in this paper, but common to all these contexts is the fact that the speaker is not primarily interested in eliciting a listener's response--in fact, no listener need be present at all. Moravcsik (1971) quotes Householder as differentiating "statements which disclaim knowledge, but exhibit indifference towards obtaining it from real questions by a feature [± Hearer] indicating hearer's involvement" (p. 81, fn. 1). We propose a similar feature though with a somewhat different definition.

As a first step, the present study was intended to assess the perceptual validity of the third category. The hypotheses were that (1) listeners can reliably identify fundamental frequency contours which display a level terminal glide rather than a terminal rise or fall, (2) listeners can reliably form a category of utterances defined by the speaker's talking to himself rather than addressing a listener, and (3) "talking-to-self" judgments, if they occur, are made of contours characterized by a moderate, even f_o, ending with a level glide.

METHOD

The stimuli were those used in a previous study (Studdert-Kennedy and Hadding, 1972, in press). They were prepared on the Haskins Laboratories Digital Spectrum Manipulator (DSM) (Cooper, 1965). This device provides a spectrographic display of a 19-channel vocoder analysis, digitized to 6 bits at 10 msec intervals, and permits the experimenter to vary the contents of each cell in the frequency-time matrix, before resynthesis by the vocoder. For the present study we were interested in the channel that displays the time course of the fundamental frequency of the utterance, since we could vary f_o by manipulating the contents of this channel.

The utterance "November"[no'vɛmbə˞] was spoken by an American male voice into the vocoder and stored in the DSM. F_o was then manipulated over a range from 85 Hz to 220 Hz. The f_o values at the most important points of the contours (starting point, peak, turning point, and end point) were chosen to represent four different f_o levels of a speaker with a range from 65 Hz to 250 Hz. The four levels were based on a previous analysis of a long sample of speech by a speaker with this particular range (Hadding-Koch, 1961:110 ff.).

The contours are schematized in Figure 1. All contours start on a f_o of 130 Hz, sustained for 170 msec, over the first syllable (the precontour). They then move, during 106 msec, to one of three peaks: 130 Hz (L, or low), 160 Hz (H, or high), 200 Hz (S, or superhigh). They proceed, during 127 msec, to one of four turning points: 100 Hz (1), 120 Hz (2), 145 Hz (3), 180 Hz (4). Finally, they proceed, during 201 msec, to one of six end points: 85 Hz (1), 100 Hz (2), 120 Hz (3), 145 Hz (4), 180 Hz (5), and 200 Hz (6). Peak, turning point, and end point are each sustained for 32 msec. The combination of three peaks, four turning points, and six end points yields 72 contours, each specified by a letter and two digits (e.g., S14 for the contour of Figure 1) and each lasting 700 msec.

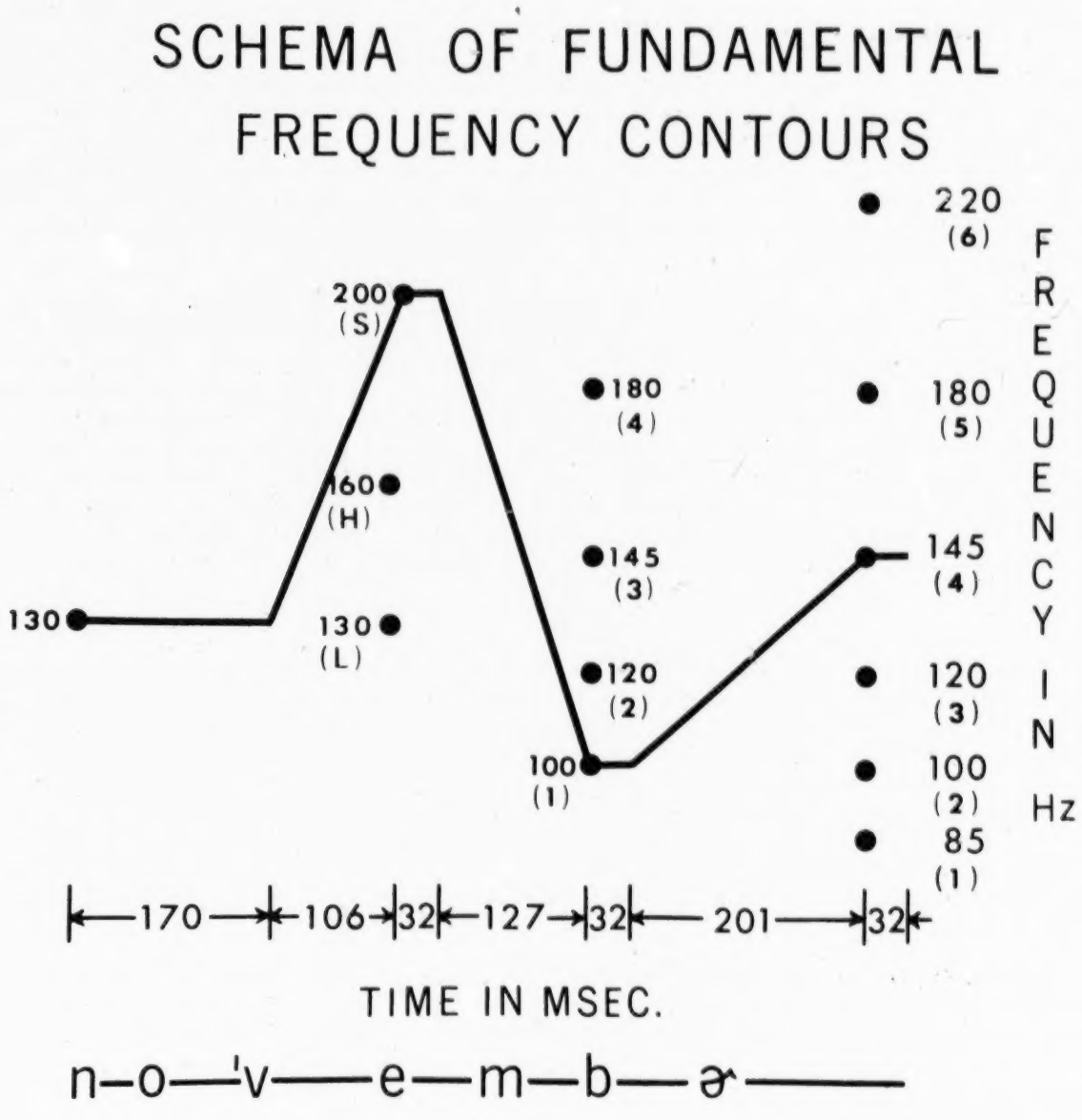


Figure 1: Schema of fundamental frequency contours imposed on the utterance "November" [no'vɛmbɚ].

The 72 contours were recorded on magnetic tape from the output of the vocoder in two forms: (1) carried on a speech wave [no'vembəʌ] and (2) as a frequency-modulated sine wave. Each set of 72 was spliced into five different random orders with a five-second interval between stimuli and a ten-second pause after every tenth stimulus.

A group of 22 Swedish graduate and undergraduate volunteers (10 of whom had served as subjects in our earlier experiments) was tested in a series of three sessions, each lasting about 45 minutes. They listened to the tests over a loudspeaker in a quiet room. In a given session they heard the five test orders for one type of stimulus only. All subjects heard the sine-wave stimuli in their first session (so as to reduce the possible influence of speech mechanisms on judgments of nonspeech stimuli). Half the group then made linguistic judgments of the speech stimuli in their second session and psychophysical judgments of the same stimuli in their third session; half the group took the tests in reverse order.

In the sine wave session and in the speech psychophysical session, subjects were asked to listen to the terminal glide of each contour and to judge whether it was rising, falling, or level in pitch. In the linguistic session, they were asked to picture three situations: a speaker addressing a question to a listener, a speaker making a statement to a listener, and a speaker not addressing a listener, but talking to himself. The subjects' task was then to listen to each utterance and assign it to its appropriate category: question, statement, or talking-to-self. The third category is not, of course, logically exclusive of the first two, and proved difficult to explain. Nonetheless, subjects agreed to try to use it and were able to do so with fair consistency.[1]

RESULTS

No systematic differences between groups due to the order in which they made their judgments were observed. Data are therefore presented for the combined groups. Figure 2 presents the sine wave, speech psychophysical, and linguistic results for the three series of contours (H3, L3, L2) in which at least one contour was judged as talking-to-self on more than 50% of the group's judgments. Percentages of fall, level, and rise judgments (sine wave and speech psychophysical) or of statement, talking-to-self, and question judgments (linguistic) are plotted against terminal glide, measured as rise (positive) or fall (negative) in Hz from turning point to end point. Each data point represents a percentage of 110 judgments (22 subjects judged each contour five times).

Consider, first, the sine-wave results (Figure 2, left column). For each series of contours the only contour judged more than 50% of the time to be terminally level in pitch is the contour for which the terminal f_0 glide was, in

---

[1] We might have avoided some of the difficulties in the linguistic session by asking subjects to use only two categories: talking to a listener and talking to self. However, we wished to compare the results with those for the psychophysical sessions, and two-category psychophysical data would have concealed potentially interesting information on the subjects' capacities for discriminating terminally level from terminally rising or falling glides.

Figure 2

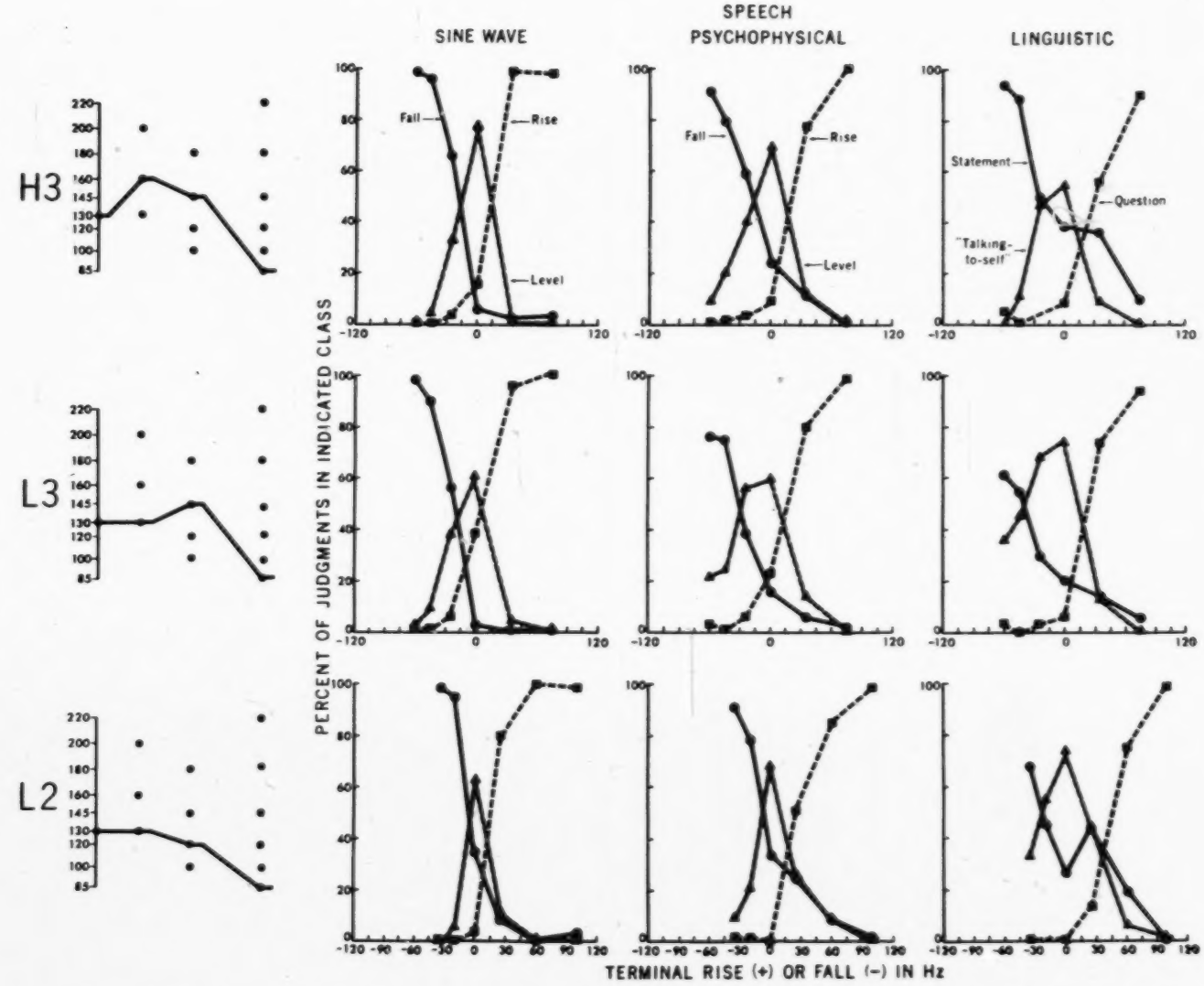


Figure 2: Percentages of fall, level, and rise responses (sine wave and speech psychophysical) or of statement, talking-to-self, and question responses (linguistic) as a function of terminal glide in Hz. Data for 22 subjects on the three series of contours for which at least one contour was judged "talking-to-self" on more than 50% of the group's judgments.

fact, level. Level judgments increase and decrease systematically on either side of this zero value, with a stronger tendency to hear a slight fall as level than a slight rise. Since level judgments never reached 100%, either for the terminally level contours of Figure 2 or for the nine other terminally level contours presented, it is evident that listeners did not find the judgment easy. However, their errors were primarily "misses" rather than "false alarms." That is to say, while four of the twelve terminally level contours failed to draw more than 50% level judgments, none of them drew as many as 50% fall or rise judgments, and none of the sixty terminally rising or falling contours drew as many as 50% level judgments.

These results are summarized in Figures 3 and 4. Figure 3 (left column) sketches the eight sine-wave contours judged level more than 50% of the time. Figure 4 (left column) sketches the four terminally level sine-wave contours for which level judgments did not reach 50%. Note that three of the latter (S12, H12, L12) display a fall from the peak to a turning point 30 Hz below the onset level of the contour; one (L45) displays a rise from the peak to a turning point 50 Hz above the onset level of the contour.

Figure 2 (center column) presents speech psychophysical results. In each graph it is again the terminally level contour that collects the highest percentage of level judgments. But the spread of level judgments over terminally falling contours is clearly broader than for the corresponding sine-wave contours. This is particularly noticeable for the L3 series, where one terminally falling contour (L33, middle row) actually draws 56% level judgments. Nonetheless, this is the only false alarm, so that, with five of the twelve terminally level contours being judged level more than 50% of the time, the errors were again primarily misses. Figures 3 (center column) and 4 (right column) summarize these results.

Figure 2 (right column) presents the linguistic judgments. In each series it is the terminally level contour that draws the highest percentage of talking-to-self judgments. But there is a clear tendency for these judgments to invade the statement category. In one series (L3, middle row) the invasion matches quite strikingly that made by level judgments into the fall category of the speech psychophysical data. However, the tendency appears in all three series so that each has a terminally falling contour that draws close to 50% talking-to-self judgments: H33 (46%), L33 (58%), L22 (55%). Figure 3 (right column) summarizes these results. Note the weight of talking-to-self judgments in the moderate to low stress peak series. No contour in the S-series meets the 50% criterion; only one in the H-series and four in the L-series meet the criterion. The L-series includes two contours with level terminal glides and two with terminal glides that fall by 20-25 Hz.

Finally, we note that, while the preferred question contours of our previous study (Studdert-Kennedy and Hadding, 1972, in press) were totally unaffected by the introduction of a third category, the preferred statement contours did not fare so well. Nine of the twenty-three statement contours on which subjects displayed at least 90% agreement in the previous study dropped below that level in the present study. Three of these (L33, L22, L23) were among the five contours collecting more than 50% talking-to-self judgments.

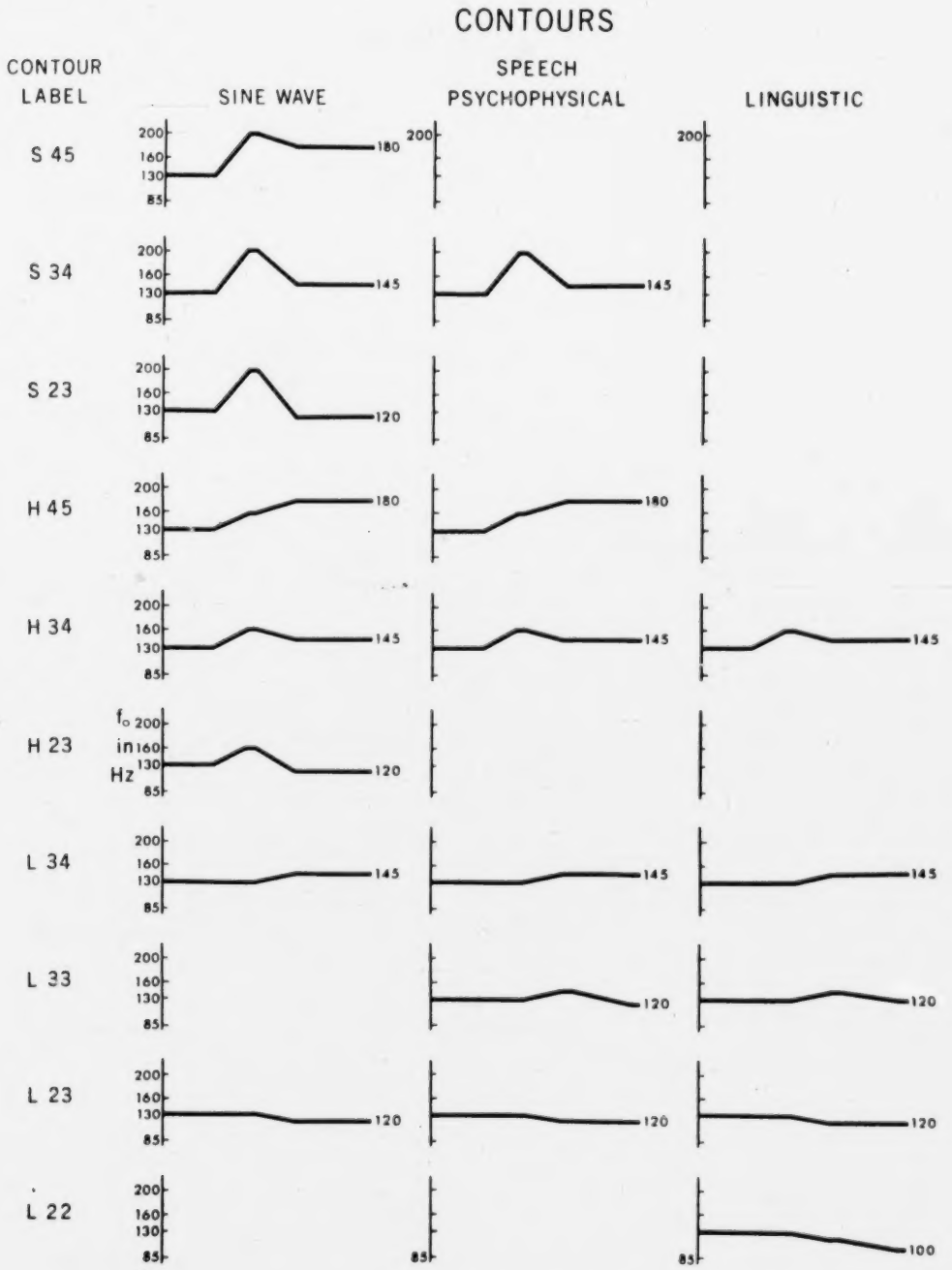


Figure 3: Schemata of all contours judged "level" (sine wave and speech psychophysical) or "talking-to-self" (linguistic) on more than 50% of the group's judgments.

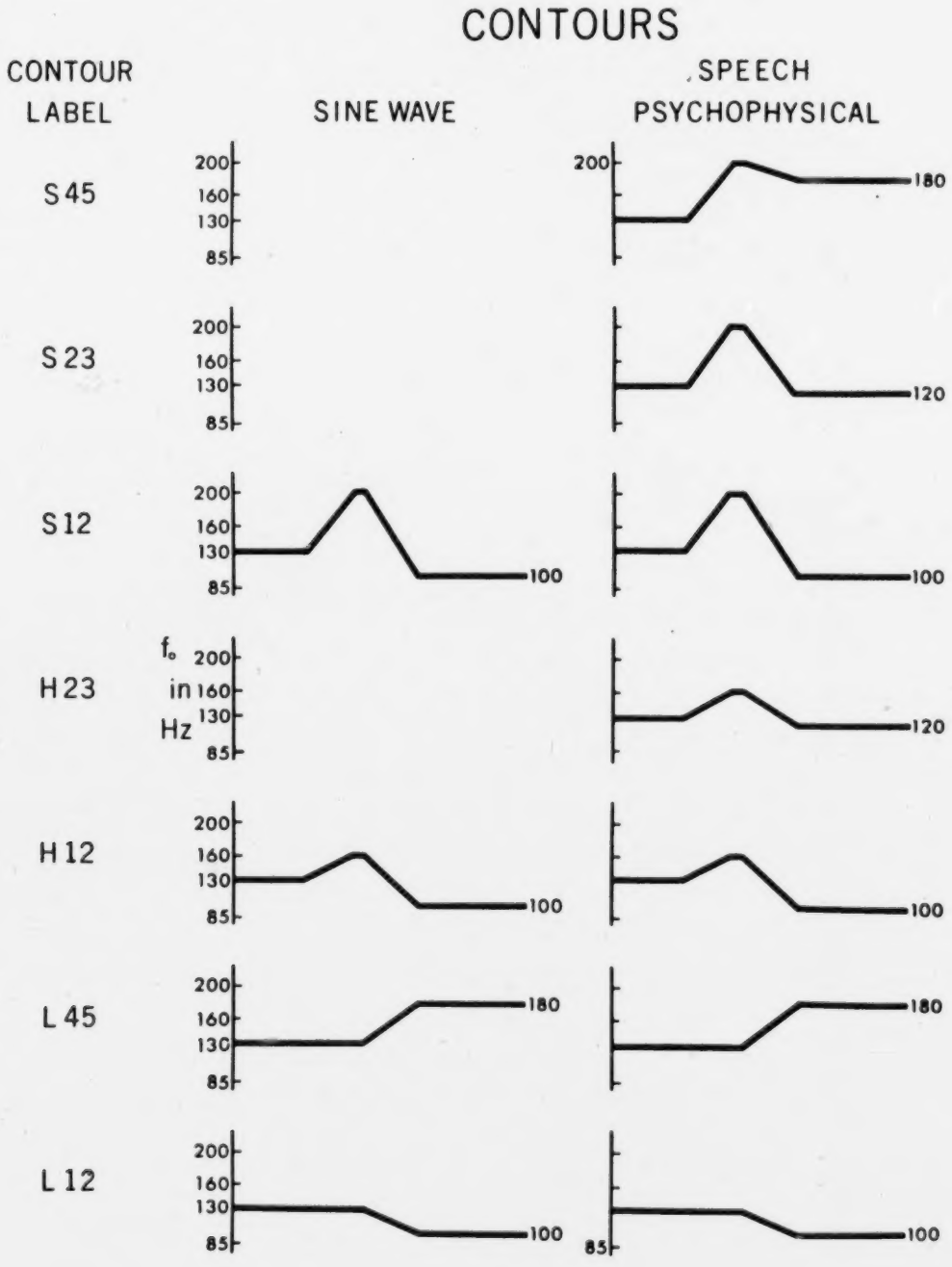


Figure 4: Schemata of all terminally level contours judged "level" on less than 50% of the group's judgments.

DISCUSSION

Listeners to brief (700 msec) frequency modulated sine-wave contours can, with some reliability, identify those that sustain a level frequency over the last 265 msec. But their performance is not perfect. While they seldom hear a rising or falling terminal glide as level, they do with fair frequency hear a level glide as rising or falling. They tend to be misled not by the initial rise to the peak, but by the rise or fall from peak to turning point, that is, by the movement of the contour during the 127 msec immediately preceding the terminal sustain: Figure 4 (left column) shows that two of the four terminally level contours that were missed display no onset-to-peak movement, but all four display a movement of at least 30 Hz from peak to turning point, and end on a frequency at least 30 Hz above or below the precontour level of 130 Hz. We may therefore say that, exactly as in our previous study (Studdert-Kennedy and Hadding, 1972, in press), listeners seem to use the precontour as an anchor and then have difficulty in separating the terminal glide from the immediately preceding section of the contour if that section displays a marked movement to a point well above or well below the anchor.

The speech psychophysical results display a similar pattern. All four of the contours missed in the sine-wave judgments were also missed in the speech psychophysical, and three more were added. Two of those added (S23, H23) display a strong fall from peak to turning point, but end on a frequency only 10 Hz below the precontour level; the other (S45) displays a fall of only 20 Hz from peak to turning point, but ends on a level 50 Hz above the precontour. In other words, there is clear overlap between sine-wave and speech psychophysical data.

Even where the two sets of data do not agree, as in the tendency for listeners to judge certain terminally falling speech wave contours as level, the errors would seem to arise from the same source as the sine wave errors, namely, from a simple inability to separate terminal glide from earlier sections of the contour. Thus, if the first 467 msec of the contour are relatively level (as in the H3 and L3 series, where all frequency variations between onset and turning point are within 30 Hz of the precontour) listeners may fail to detect the slight terminal fall and then judge it to be level (see Figure 2, center column). In other words, they do not, as might be predicted from the analysis-by-synthesis model of Lieberman (1967), accept a falling glide as level when the stress peak is exceptionally high, but rather when the entire section of the contour preceding the terminal glide is relatively low and level.

Turning to the linguistic data, we may say that listeners are indeed able to identify an utterance as that of a speaker talking to himself and that they may even do so with more consistency than they make the corresponding psychophysical judgment (see Figure 2: L34, L23). Nonetheless, they are not perfectly consistent. One reason for this is that the categories statement and talking-to-self are not mutually exclusive, and so compete for certain contours. This is evidenced by the tendency of talking-to-self judgments to take over the statement category at the level terminal glide and by the fact that three of the four contours collecting more than 50% talking-to-self judgments in the present study drew more than 90% statement judgments in our previous study. Combined with this, a second factor may have contributed to listener uncertainty: intensity. Talking to ourselves we speak softly. But all utterances in the present study were of equal intensity so that a listener, choosing between the two competing

categories, may have been pushed toward statement by a relative intensity more apt for addressing others than self.

In considering the linguistic results, we should bear in mind that, while psychophysical judgments were made on the terminal glide, linguistic judgments were made on the entire contour. If, therefore, sine wave, speech psychophysical, and linguistic judgments coincide, we may reasonably conclude that terminal glide controlled linguistic decision.[2] From Figures 2 and 3 it is evident that, as far as the third category (level or talking-to-self) is concerned, the three groups of judgments do coincide on certain contours that exhibit a level (H34, L34, L23) or slightly falling (H33, L33) terminal glide. This agreement confirms the importance of the terminal glide in linguistic judgments of intonation contours. While our previous study gave clear evidence of the connection between terminal rise/fall and judgments of question/statement, the present study demonstrates a clear connection between terminal sustain and judgments of talking-to-self.

However, terminal glide is not the only determinant of linguistic decision. Figure 3 shows that one terminally falling contour (L22) was judged as talking-to-self more than 50% of the time, but did not reach criterion on speech wave level judgments, while two speech wave contours (S34, H45), correctly heard as level more than 50% of the time did not reach criterion on talking-to-self judgments. In fact, of the five acceptable talking-to-self contours, four display no stress peak (L34, L33, L23, L22), and one displays a moderate peak but then drops to within 15 Hz of the precontour level (H34). Evidently, we expect people talking to themselves not only to end their utterances with a level (or slightly falling) glide, but also to maintain an even, low to moderate pitch over earlier sections of the contour. The initial hypothesis is therefore largely confirmed.

To sum up, this study has provided experimental support for the validity of the third category described by Hadding-Koch (1961), and for the adoption of a new prosodic feature, [± Listener], implemented by variations in fundamental frequency and, perhaps, intensity. The communicative function of the feature is presumably to draw and hold a listener's attention.

REFERENCES

Cooper, F. S. (1965) Instrumental methods for research in phonetics. Proceedings of the Vth International Congress of Phonetic Sciences, Münster 1964, (Basel: S. Karger) 142-171.

Hadding-Koch, K. (1961) Acoustico-phonetic Studies in the Intonation of Southern Swedish. (Lund: Gleerups).

Hadding-Koch, K. and M. Studdert-Kennedy. (1964) An experimental study of some intonation contours. Phonetica 11, 175-185.

---

[2] Where speech psychophysical judgments follow the linguistic more closely than the sine-wave judgments, it is also possible, as remarked in our previous study, that linguistic decision controlled auditory judgment rather than vice versa. Obvious candidates for this interpretation are contours of the L3 series (Figure 2, middle column). However, it is difficult to understand why this hypothetical process should have come into play for this series, but not for others.

Hadding-Koch, K. and M. Studdert-Kennedy. (1965a) Intonation contours evaluated by American and Swedish test subjects. Proceedings of the Vth International Congress of Phonetic Sciences, Münster 1964, (Basel: S. Karger) 326-331.

Hadding-Koch, K. and M. Studdert-Kennedy. (1965b) A study of semantic and psychophysical test responses to controlled variations in fundamental frequency. Studia Linguistica XVII, 65-76.

Lieberman, P. (1967) Intonation, Perception and Language. (Cambridge, Mass.: MIT Press).

Moravcsik, E. A. (1971) Some crosslinguistic generalizations about yes-no questions and their answers. Working Papers on Language Universals (Stanford University) 7, 45-181.

Studdert-Kennedy, M. and K. Hadding. (1972) Further experimental studies of f_o contours. Proceedings of the VIIth International Congress Phonetic Sciences, Montreal, 1971.

Studdert-Kennedy, M. and K. Hadding. (in press) Auditory and linguistic processes in the perception of intonation contours. Language and Speech. (Also in Lund University Working Papers 5, 1971 and in Haskins Laboratories Status Report on Speech Research SR-27, 1971.)

Uldall, E. T. (1962) Ambiguity: question or statement? or "Are you asking me or telling me?" Proceedings of the IVth International Congress of Phonetic Sciences, Helsinki 1962, (The Hague: Mouton) 779-783.

Discrimination of Intensity Differences on Formant Transitions In and Out of Syllable Context

M. F. Dorman[+]
Haskins Laboratories, New Haven, Conn.

According to traditional psychoacoustics, listeners can discriminate among many more acoustic stimuli along nonspeech continua than they can identify absolutely. For example, listeners can discriminate about 1200 different pitches, yet can identify absolutely only about 7 ±2 (Miller, 1956; Pollack, 1952).

This type of relationship between identification and discrimination does not hold, however, for certain acoustic stimuli, the speech sounds stop consonants. In discrimination tests, when listeners are presented with synthetic speech stimuli which differ in extent and direction along the acoustic continuum of the second formant transition, discrimination is essentially perfect between stimuli drawn from different phonetic categories (e.g., /ba-da/), but is near chance for physically different stimuli drawn from the same phonetic category. Thus, the discrimination of differences in extent and direction of the second formant transition (also for the cutback of the first formant, the cue for manner of articulation) is little better than an absolute phonetic categorization (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Pisoni, 1971). Liberman and his colleagues have termed this type of relationship between identification and discrimination, which is apparently unique to the perception of speech, categorical perception (Mattingly, Liberman, Syrdal, and Halwes, 1971; Studdert-Kennedy, Liberman, Harris, and Cooper, 1970).

The categorical perception of stop consonants has been interpreted as demonstrating that after recoding of the acoustic signal into a phonetic representation, auditory information cannot be readily retrieved from short term auditory memory (Pisoni, 1971). From the studies cited above, however, it is not clear whether all of the acoustic information on the formant transitions suffers a fate in short term auditory memory similar to the acoustic information which directly cues different phonetic categories. For example, is intensity of the formant transitions (information which does not directly signal phonetic categories) as poorly discriminated within a phonetic category as changes in the extent and direction of the second formant transition of first formant cutback? The purpose of the present experiment was to determine the discriminability of differences in intensity of formant transitions on a computer generated stop consonant-vowel syllable /bae/.

[+]Also Herbert H. Lehman College of the City University of New York.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

METHOD

Subjects. Ten adults, who had previously participated in research with synthetic speech, served as Ss.

Preparation of stimuli. The stimuli were generated on the Haskins Laboratories computer controlled parallel resonance synthesizer.

One set of stimuli was created by first synthesizing a three formant, stop consonant-vowel syllable /bae/. For the first stimulus in this set, the overall intensity of the first 50 msec (the duration of the formant transitions) was the same as the remaining 200 msec of the syllable. This will be referred to as a Δ0 db stimulus. For the second and third stimuli in this set the first 50 msec was 7.5 db and 9 db respectively less intense than the initial portion of Δ0 db stimulus. The following 200 msec of both stimuli was the same intensity as the Δ0 db stimulus.

A second set of stimuli was created by simply eliminating the vowel steady-state formants from the three /bae/ syllables, leaving the formant transitions isolated (see Figure 1). The resulting three stimuli no longer sounded like speech, but rather nonspeech "chirps" or pitch glides.

A third set of stimuli was created by synthesizing a 250 msec steady-state vowel /ae/ (see Figure 1). For the first stimulus in this set the intensity of the initial 50 msec was as intense as the remaining 200 msec. For the second and third stimuli in this set, the initial 50 msec was 7.5 db and 9 db respectively less intense than the initial portion of the Δ0 db stimulus. The following 200 msec of both stimuli was the same intensity as the Δ0 db stimulus.

With the aid of the computer controlled synthesizer, three test sequences were recorded on audio tape. One sequence contained the stop consonant-vowel stimuli; a second, the isolated formant transition stimuli; and the third, the steady-state vowel stimuli. The stimuli in all of the sequences were arranged in pairs. The first member of each pair was a Δ0 db stimulus, the second member either a Δ0 db, Δ7.5 db, or Δ9 db stimulus. Each test sequence contained a randomized sequence of 18 pairs of Δ0 db-Δ0 db stimuli, 18 pairs of Δ0 db-Δ7.5 db stimuli, and 18 pairs of Δ0 db-Δ9 db stimuli. The interstimulus interval within stimulus pairs was 500 msec. The interval between successive pairs was 4 sec.

Procedure. Each S was tested individually in a sound attenuated room. The stimuli were presented via TDH-39 headphones at 72 db SPL (for the Δ0 db stimuli). Each S heard all three test sequences. The order of the sequences was partially randomized across Ss. The Ss were informed as to the type of sounds they would hear and were instructed to listen for differences in intensity of the initial portions of the stimulus pairs. Ss wrote either same (S) or different (D) on printed answer sheets. Before each test sequence, Ss were presented 18 practice trials of stimulus pairs and were concurrently shown the correct responses on an answer sheet.

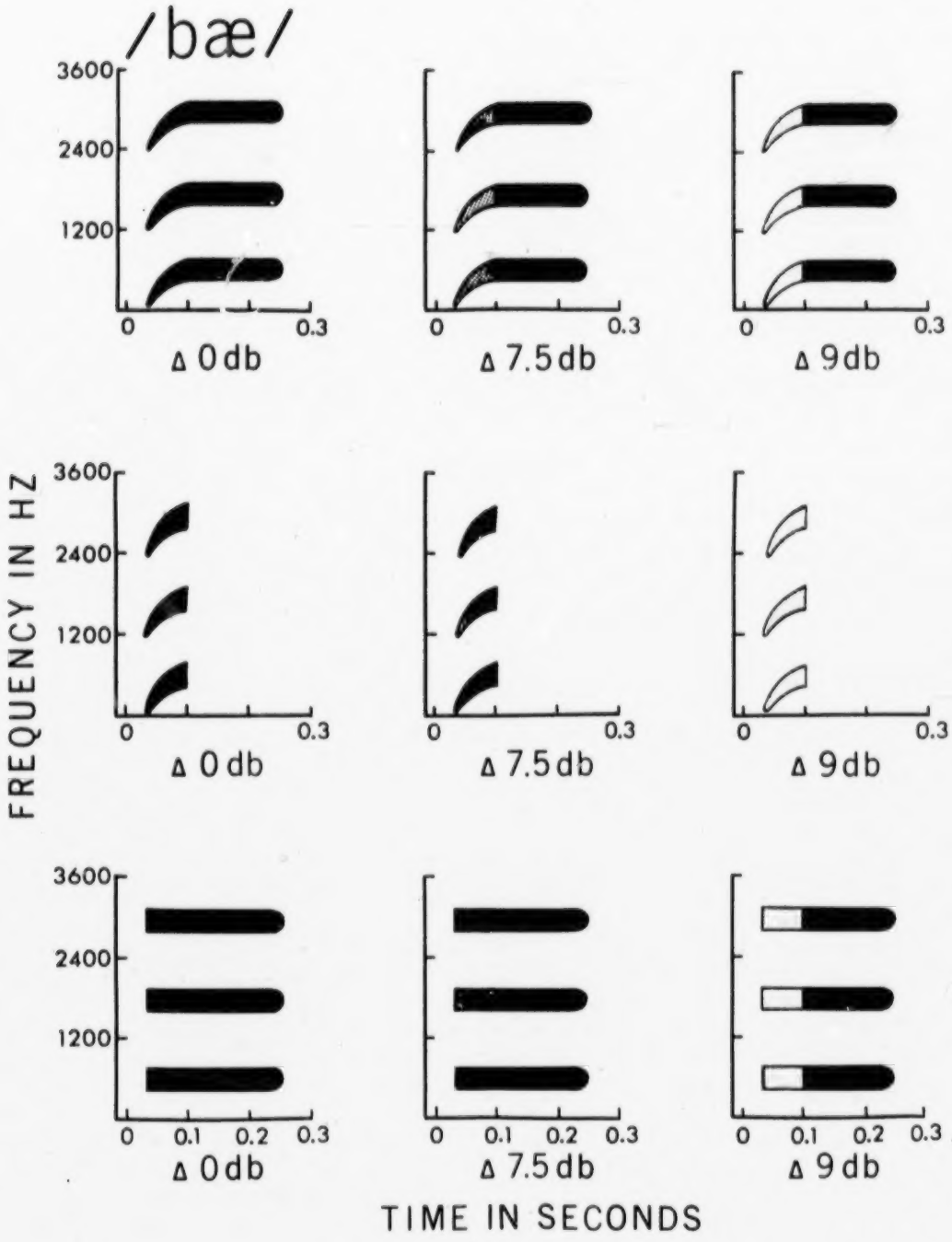


Figure 1: Schematic spectrographic patterns for /bae/, the isolated formant transitions from /bae/, and /ae/. For each stimulus type the initial 50 msec was either the same intensity as the remaining 200 msec, or was 7.5 db or 9.0 db less intense.

RESULTS

The probability of a "different" response when the stimulus pairs were in fact different (i.e., P("D"/D)] for the three experimental conditions is shown in Figure 2. The Δ0 db-Δ7.5 db and Δ0 db-Δ9 db stimulus pairs, in the syllable context /bae/, were discriminated correctly on 42% and 54% of the trials respectively. These same intensity differences in the isolated formant context were discriminated correctly on 96% and 100% of the trials respectively. The intensity differences in the steady state vowel context were discriminated correctly on 92% and 98% of the trials respectively.

By scoring the "catch" trials, that is the Δ0 db-Δ0 db stimulus pairs, in all conditions, the detectability (d') of the intensity differences independent of response bias was computed. In the syllable context, Ss averaged 30% false alarms; in the isolated formant context, 7% false alarms; and in the steady state vowel context, 3% false alarms. Taking these false alarm rates into account, the d' scores for the discriminability of the 7.5 db and 9.0 db intensity differences in the syllable context were .32 and .62; in the isolated formant context, 3.72 and >3.79; in the steady-state vowel context, 3.28 and 3.93.

DISCUSSION

The 7.5 db and 9 db differences in intensity of the formant transitions, when in the syllable context /bae/, were discriminated essentially at chance. However, the 30% false alarm rate indicates a somewhat conservative criterion for accepting a difference in intensity. This suggests that some, although very little, information about the intensity of the formant transitions was available for recall from short term auditory memory. This outcome is consistant with the previously cited studies on the discriminability of within-category acoustic information, such as extent and direction of the second formant transition, which directly cues different phonetic categories (Mattingly et al., 1971; Pisoni, 1971). The poor discrimination was not due to the intensity differences being in themselves not discriminable, since in the isolated formant context, the differences were discriminated essentially perfectly.

It might be argued, however, that the poor discrimination of the intensity differences on the formant transitions was due to backward masking and not due to processes unique to the coding of speech. That is, the steady state vowel formants, because they follow and are more intense than the formant transitions, mask some of the acoustic information on the formant transitions. This interpretation can be assessed by examining the discriminability of the 7.5 db and 9 db intensity differences in the steady-state vowel context. Backward masking ought to have been maximized in this condition since the formant frequencies for the initial 50 msec and the following 200 msec were identical.

As shown in Figure 2, however, Ss discriminated intensity differences in the steady-state vowel context about as well as in the isolated formant context. It does not seem likely then, that the differences in discriminability of the intensity of the formant transitions, in and out of syllable context, was a function of backward masking.

We may account for these data by assuming that after the acoustic cues for stop consonant formant transitions have been recoded into a phonetic representation,

Figure 2

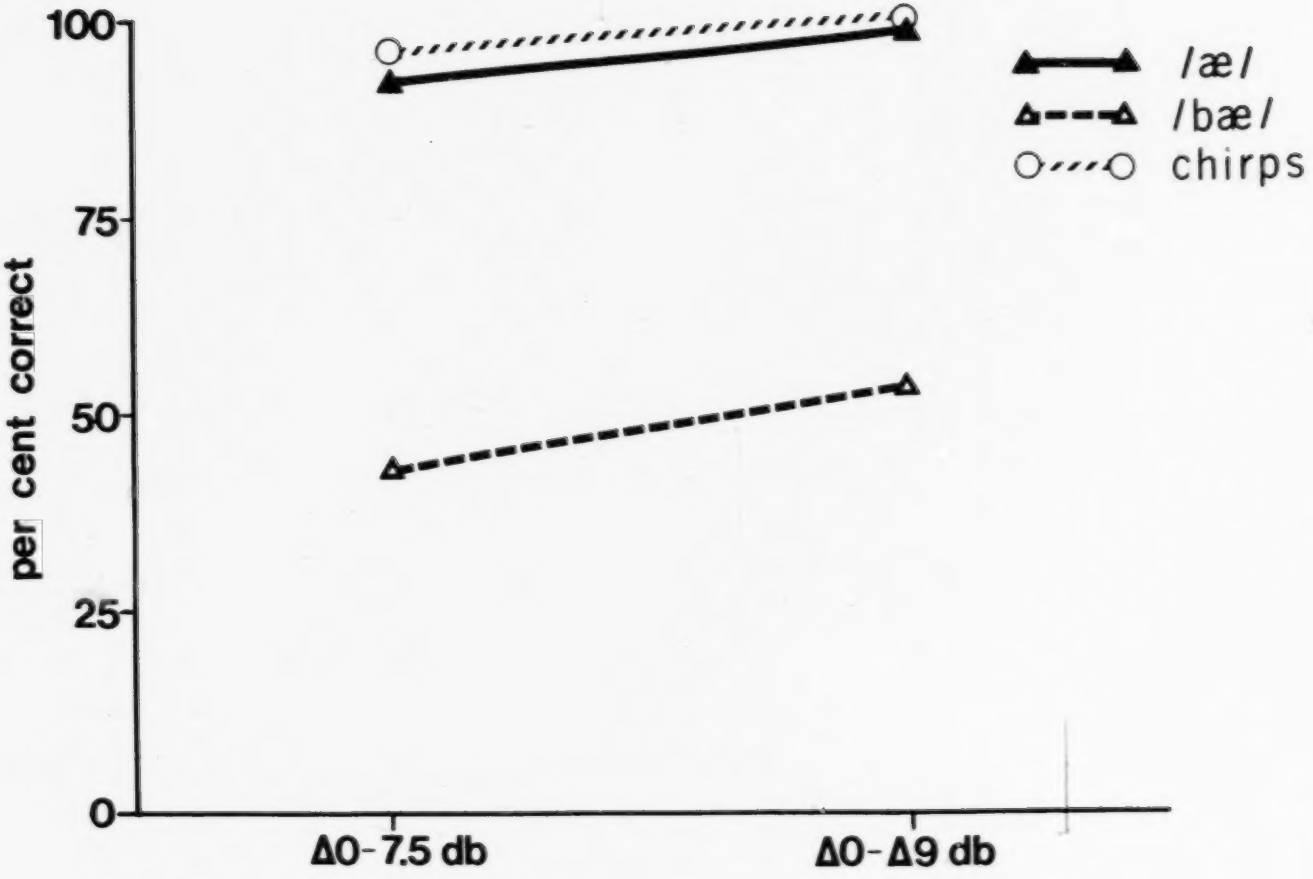


Figure 2: P("D"/D) for the 7.5 db and 9 db differences in intensity of formant transitions in syllable context, in isolation, and for steady-state vowels.

all of the acoustic information is stored in a relatively inaccessible short term auditory memory. All auditory information should be then equally inaccessible for recall. This would account for the poor discrimination within a stop consonant category of differences in first and second formant transitions, as well as the poor discrimination of intensity differences on formant transitions. On this view it would be expected that stop consonant vowel syllables from the same phonetic category, which differed only in the pitch or duration of the formant transitions, would also be discriminated just slightly better than chance.

REFERENCES

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Miller, G. A. (1956) The magical number seven, plus or minus two, or some limits on our capacity for processing information. Psychol. Rev. 63, 81-96.

Pisoni, D. (1971) On the nature of categorical perception of speech sounds. Ph.D. thesis, University of Michigan. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Pollack, I. (1952) The information in elementary auditory displays. J. Acoust. Soc. Amer. 24, 745-749.

Studdert-Kennedy, M., A. M. Liberman, K. S. Harris, and F. S. Cooper. (1970) The motor theory of speech perception: a reply to Lane's critical review. Psychol. Rev. 77, 234-249.

Phonological Fusion in Synthetic and Natural Speech

James E. Cutting[+]
Haskins Laboratories, New Haven, Conn.

Synthetic speech sounds artificial, at least using present speech synthesis systems. This is a problem which has nagged not only those who use synthetic speech as an applied tool, as in reading machines for the blind (Cooper, Gaitenby, Mattingly, Nye, and Sholes, 1972), but also those who use synthetic speech as a research tool. What kind of perceptual artifacts are inherent to the "cold-in-the-head" quality of synthetic speech? Happily they appear to be minimal. In tasks involving memory span for specific orders of digits, synthetic and natural speech stimuli yield the same pattern of results (Crowder, 1971). In dichotic listening tasks involving the perceptual rivalry of two speech signals, the same direction and order of magnitude in ear advantages is found (Shankweiler and Studdert-Kennedy, 1967; Studdert-Kennedy and Shankweiler, 1970). Is this also true for phonological fusion? Day (1968) used natural speech stimuli, and Cutting (in preparation) used synthetic speech stimuli to explore aspects of phonological fusion. To what extent are the two situations comparable?

METHOD

Stimuli. Four sets of stimuli of the same general pattern were selected: the PAY set (PAY, RAY, LAY); the BED set (BED, RED, LED); the CAM set (CAM, RAM, LAMB); and the GO set (GO, ROW, LOW). All stimuli within each set were identical except for the initial phoneme. Two versions of each stimulus were prepared, one synthetic and one natural speech. The synthetic stimuli were prepared on the Haskins Laboratories parallel resonance synthesizer and were the same as those used by Cutting and Day (1972) in a previous dichotic fusion study. All stimuli within each set of synthetic stimuli were identical in pitch, intensity, and duration, and all had the same acoustic structure except for the initial 150 msec. Liquid stimuli within each set differed only in the direction and extent of the third formant (F3) transition. This acoustic cue is the primary cue for discriminating the two liquids, /r/ and /l/ (O'Connor, Gerstman, Liberman, Delattre, and Cooper, 1957). Natural speech versions of the same items were spoken by the author and recorded on audio tape. Although an effort was made to match the utterances in terms of pitch, intensity, and duration, some variation was unavoidable. Both synthetic and natural speech stimuli were then digitized and stored on disc file for the preparation of diotic and dichotic tapes. Figure 1 shows synthetic and natural speech versions of PAY and LAY. Notice that the synthetic stimuli are much "cleaner" signals than the natural speech. The bands of high

---

[+]Also Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

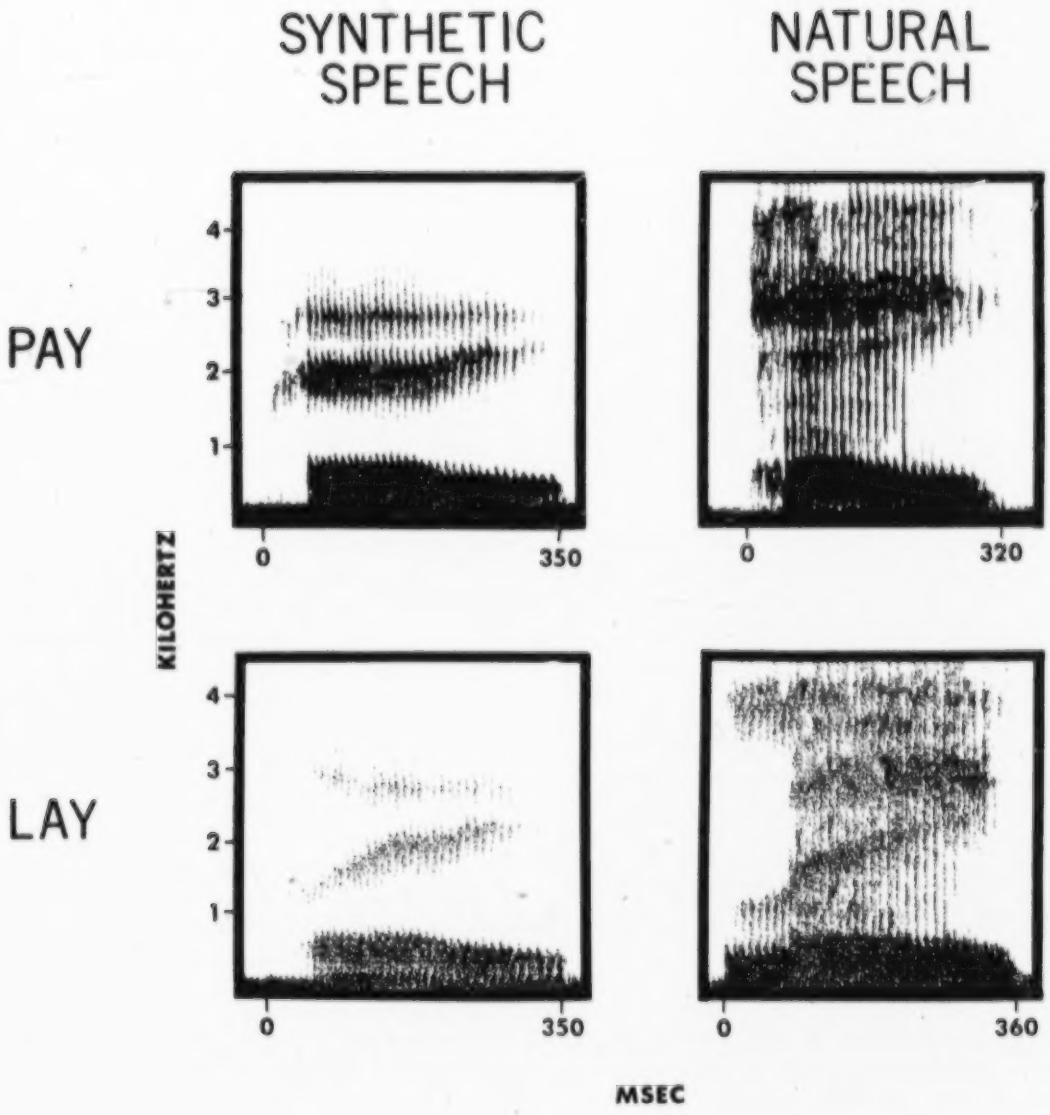


Figure 1: Spectrograms of synthetic and natural speech versions of PAY and LAY.

energy (formants) are clearly distinct from the areas of low energy in the synthetic stimuli, whereas the energy is more irregularly smeared across the spectrum in the natural speech.

Subjects. Sixteen Yale University undergraduates participated in two tasks: identification and fusion.

Task 1: Identification

Stimuli, tapes, and procedure. All stimuli, natural and synthetic, were recorded on a diotic identification tape. The tape consisted of a random sequence of 120 trials: (4 sets of stimuli) x (2 versions of each set: natural and synthetic) x (3 stimuli per set) x (5 observations per stimulus). Subjects wrote down the entire word that they heard presented. There was a 3 second interval between stimuli.

Results. All stimuli were highly identifiable. Synthetic and natural speech stimuli were correctly identified on better than 93 percent of all trials. Cluster responses (such as PLAY) were reported on less than 2 percent of all trials.

Task 2: Fusion

Stimuli and tapes. All stimuli used in the identification task were also used in the fusion task. However, instead of presenting one stimulus at a time, two stimuli were presented, one to each ear. The dichotic pair consisted of a stop stimulus and a liquid stimulus from the same set, for example PAY/LAY or PAY/RAY. Figure 2 shows the stimuli and the possible fusions that could occur. Note that all stimuli and possible fusions are monosyllabic, high frequency English words. Most have a Thorndike and Lorge (1944) frequency of at least 100 per million. Possible fusions of stop + /r/ and stop + /l/ within each set occur with approximately equal frequency. All stop stimuli cluster with both /r/ and /l/ in initial position, /p/ in PAY, /b/ in BED, /k/ in CAM, and /g/ in GO. No alveolar stop stimuli were chosen because /tl/ and /dl/ clusters do not appear in initial position in English. Both front vowels (/ɛI/ in PAY, /ɛ/ in BED, and /ae/ in CAM) and a back vowel (/oʊ/ in GO) occur in the stimuli.

Two tapes were prepared, one using the synthetic stimuli and one using natural speech stimuli. Three lead times were selected: the stop-initial stimulus could begin 50 msec before the liquid-initial stimuli, they could begin simultaneously, or the liquid could begin 50 msec before the stop. There were 96 dichotic items per tape: (4 sets of fusible stimuli) x (2 liquids per set) x (3 lead times) x (2 channel arrangement per pair) x (2 observations per dichotic item). All subjects listened to both types of stimuli: half listened first to the synthetic speech pairs and then to the natural speech pairs, while the other half listened in the reverse order.

RESULTS AND DISCUSSION

Overall fusion rate. Fusion responses occurred readily for all stimuli. Fusion rates, however, were higher for synthetic stimuli than for natural speech stimuli. Subjects fused synthetic pairs on 61 percent of all trials, while the corresponding fusion rate for the natural speech pairs was only 31 percent. This 2/1 ratio in fusion level was highly significant: all 16 subjects showed results in this direction ($z = 3.75$, $p < .0001$).

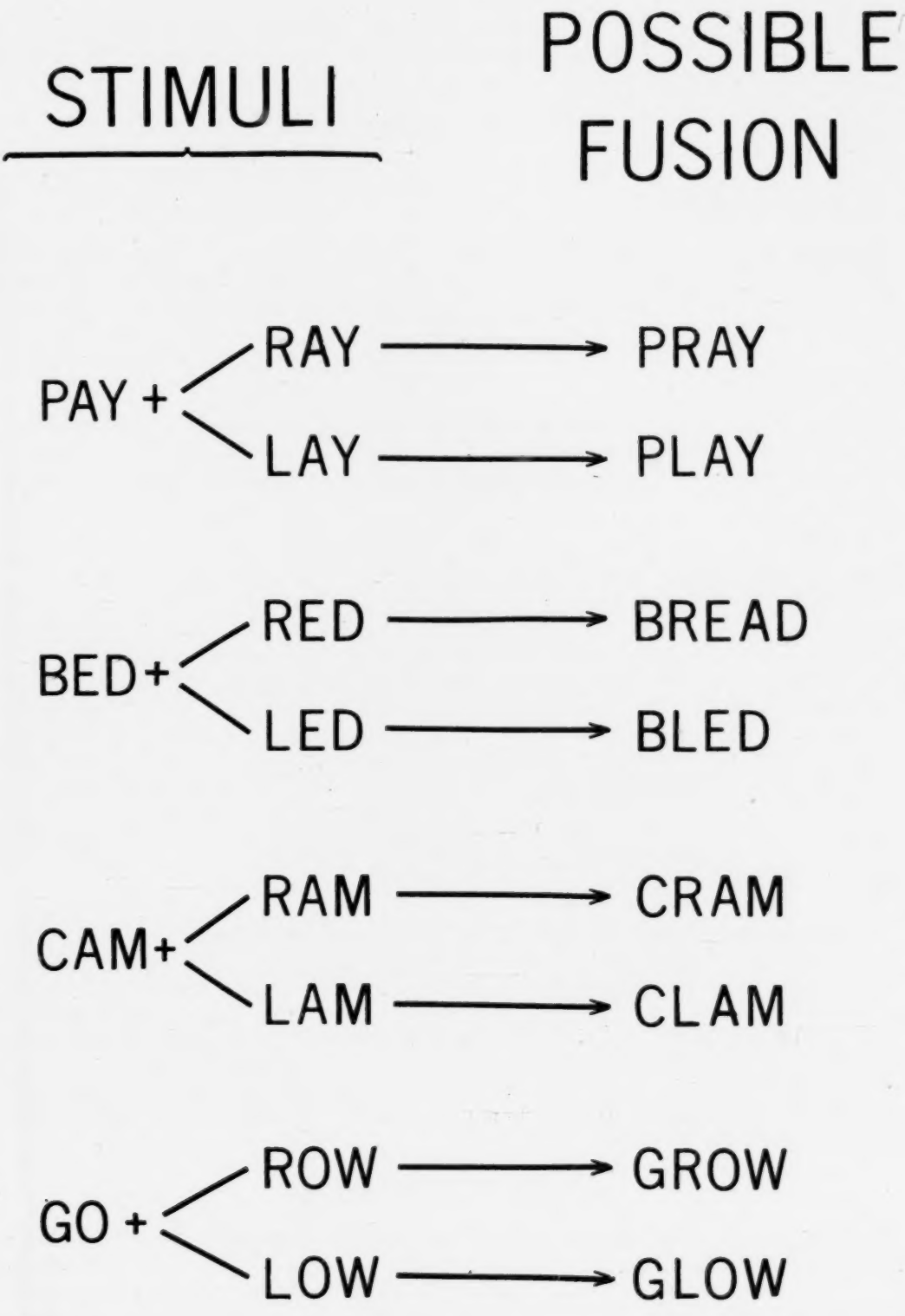


Figure 2: Stimuli and possible fusions in the fusion task.

The explanation for the differential fusion rates may lie in the acoustic compatibility of the stimuli. Synthetic fusible pairs were made to be identical with respect to pitch, intensity, and duration, whereas the natural speech pairs varied somewhat in each of these parameters. Natural speech versions were quite comparable to one another, but differed by as much as 14 Hz in pitch, 2 db in intensity, and 40 msec in duration. As shown in Figure 1, the stimuli PAY and LAY are much more similar in the synthetic renditions than in the natural speech. The effects of pitch and intensity differences in the stimuli have been explored experimentally by Cutting (in preparation) and they were found to have no effect on fusion rate. Thus, the difference in fusibility must be accounted for on other grounds. Perhaps the difference in fusion rate is caused by the general clarity of the signals: synthetic versions are very clear and spare stimuli, whereas the natural speech versions are full of noise and general formant irregularities. The order in which the subjects listened to the synthetic and natural speech stimuli was not a significant factor.

Fusion of stop + /r/ and stop + /l/ stimuli. Figure 3 shows the fusion rates for stop + /r/ and stop + /l/ stimuli in both synthetic and natural speech versions. In general, fusion rates were slightly higher for stop + /l/ pairs than for stop + /r/ pairs (e.g., PAY/LAY versus PAY/RAY). This difference was not significant for the synthetic stimuli: the respective fusion rates for stop + /r/ and stop + /l/ pairs was 58 and 64 percent. The difference, however, was significant for the natural speech stimuli: the fusion rates for stop + /r/ and stop + /l/ were 26 and 36 percent respectively. Thirteen of 16 subjects showed this pattern, a result which replicates Day (1968). She found significant differences in fusion rates for the two types of pairs, using natural speech disyllabic pairs such as BANKET/LANKET and PAHDUCT/RAHDUCT. In the present study the differences in fusion rates for stop + /r/ and stop + /l/ pairs is in the same direction for both synthetic and natural speech pairs, but the proportional difference is much greater for the natural speech.

How might we explain the difference between the fusibility of /l/ and /r/ stimuli in synthetic and natural speech versions? Synthetic versions of liquid stimuli such as RAY and LAY differed only in the direction and extent of the F3 transition, a very small but phonetically important difference. Natural speech versions of the same items differed along other dimensions as well. Although the F3 transition is the primary cue which distinguishes the two liquids, other secondary cues do exist in naturally produced liquids. Duration of the formant transitions, amount of steady-state voicing preceding the transitions, and the abruptness of the overall intensity rise time at the beginning of the release of the liquid are among the differences of secondary importance. Any of these differences may have facilitated the differential fusion rates in the two types of natural speech pairs.

Misperceptions. Figure 3 shows that misperceptions occurred quite frequently for both synthetic and natural speech pairs. The most common misperception was a stop + /l/ fusion response for a stop + /r/ stimulus pair, for example PAY/RAY⟶ PLAY.[1] While /l/ was substituted for /r/ quite frequently, the reverse substitution was much less frequent. Day (1968) and Cutting and Day (1972) have reported

---

[1]The arrow should be read as "yields."

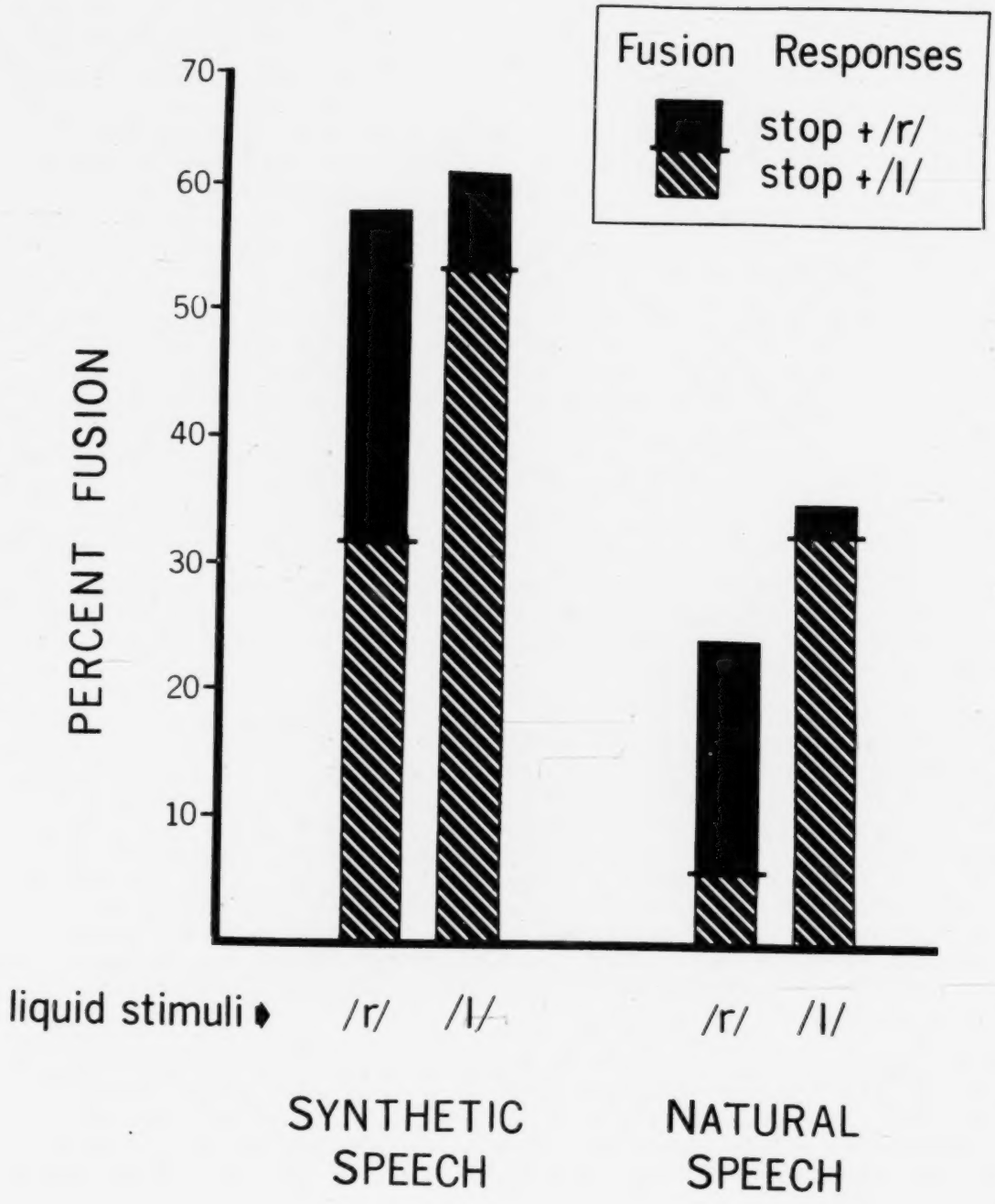


Figure 3: Results of the fusion task.

this /l/-for-/r/ substitution phenomenon. It cannot be accounted for by the frequency of these clusters in English. In fact, frequency data show the reverse trend: stop + /r/ clusters outnumber stop + /l/ clusters in English by a ratio of nearly 2/1 (Day, 1968). In the present study /l/-for-/r/ substitutions occurred for both types of stimuli at approximately the same rate. For comparison, we can make a ratio of misperceptions--misperception being defined as a fusion response which contains a liquid different than the one presented. In the numerator is the percentage of /l/-for-/r/ substitutions in fusion responses (e.g., PAY/RAY⟶PLAY) and in the denominator is the percentage of the reverse substitution (e.g., PAY/LAY⟶PRAY). For synthetic speech pairs this ratio was 3.3/1, while for natural speech pairs it was 3.4/1. The difference between the two ratios is not significant, and shows rather convincingly that the /l/-for-/r/ substitutions are not an artifact of synthetic speech but are inherent to the phonological fusion task. Furthermore, they occurred at approximately the same rate for all stimulus sets.

Fusion for different sets of stimuli. Fusion rates differed across sets of stimuli. Table 1 shows the fusion rates for the four sets of stimuli in both synthetic and natural speech versions. Fusion rates for the PAY and GO sets were consistently higher than those for the BED and CAM sets. One explanation for these differences may relate to the frequency of occurrence of these fused

TABLE 1: Percent fusion for each set in both synthetic and natural speech versions.

SET	PERCENT FUSION		
	Synthetic	Natural	Mean
PAY set	74	37	55
BED set	47	26	37
CAM set	60	23	42
GO set	64	38	51
Mean	61	31	46

responses as words in English: according to Thorndike and Lorge (1944) and Carroll, Davies, and Richman (1971) the words PRAY, PLAY, GROW, and GLOW occur considerably more frequently in general publications that BREAD, BLED, CRAM, and CLAM. Day (1968) has shown that fusions occur more frequently when the fused outcome is an acceptable English word than when it is not (although nonword fusions do occur, for example, GORIGIN/LORIGIN⟶GLORIGIN). Perhaps within the word category, frequency of occurrence in the natural language also plays a role.

Phonetic differences among the stimulus sets cannot account for the differential fusion rates of each set. Place differences among the stop-initial stimuli were not correlated with fusion rate: /p/ in PAY and /b/ in BED are both labials, and yet these sets showed appreciably different fusion rates. The same relationship occurred for /k/ in CAM and /g/ in GO: both are velars, and yet large

differences in fusion rate were observed between the two sets. The voicing dimension fares little better as a predictor of fusion rates: /p/ in PAY and /k/ in CAM are both voiceless, yet one set (PAY) fused at a significantly higher rate than the other (CAM). Again, the same relation occurs for /b/ in BED and /g/ in GO; both stops are voiced and yet the sets were quite different in their fusion rate. Vowel distinctions are also little help. The /ɛI, ɛ, ae/ in PAY, BED, and CAM are all front vowels, while /oʊ/ in GO is a back vowel. No pattern is evident here for fusion rate according to vowels in the stimulus sets: the set with the back vowel (GO) and one set with a front vowel (PAY) fused at comparatively high levels, while the other two sets with front vowels (BED and CAM) fused at a much lower level. Thus, phonetic differences appear to have less to offer in the prediction of fusion rate than does frequency of occurrence.

Ear effects. There were none. Fusion occurred equally readily when PAY was presented to the right ear and LAY to the left as in the reverse condition. Ear differences are the result of the competition of linguistic information. In phonological fusion, there is no competition and there is no information loss: if the stimuli are PAY/LAY and the subject reports hearing PLAY, he has reported all of the linguistic information presented to him, merely reorganized into a perceptual whole. Without the loss of information there can be no decrement in performance, and without the decrement there is no ear effect. In the present study when the subject did not fuse, he invariably reported hearing only the stop stimulus (PAY/LAY⟶PAY). Since the stop stimulus appeared equally often on either channel, there was no ear advantage for the nonfused trials either.

There were no significant channel effects, earphone effects, or any other procedural effects. For a discussion of individual differences in fusion rates and the effect of lead times on fusion rate see Cutting (in preparation).

CONCLUSION

While there were differences in fusion rate between synthetic and natural speech stimuli, the rules which govern their fusibility are comparable. This conclusion is based on information from two sources. First, the pattern of fusibility for the four sets of stimuli were quite similar for both the natural and synthetic speech stimuli. In both cases the PAY and GO sets fused more readily than the CAM and BED sets. Second, the pattern of misperceptions for the two types of stimuli was almost identical. The phoneme /l/ was quite frequently substituted for /r/ on a fusion response; for example, PAY/RAY⟶PLAY. The reverse substitution was much less common. The ratios for the two types of misperceptions were comparable for the synthetic and natural speech stimuli. Thus, one may generalize about aspects of phonological fusion from the results of tasks using synthetic stimuli without fear of artifacts inherent in synthetic speech.

REFERENCES

Carroll, J. B., P. Davies, and B. Richman, eds. (1971) Word Frequency Book. (New York: Houghton Mifflin Co.).

Cooper, F. S., J. H. Gaitenby, I. G. Mattingly, P. W. Nye, and G. N. Sholes. (1972) Audible outputs of reading machines for the blind. Haskins Laboratories Status Report on Speech Research SR-29/30, 91-95.

Crowder, R. G. (1971) Waiting for the stimulus suffix: decay, delay, rhythm, and readout in immediate memory. Quart. J. Exp. Psychol 23, 324-340.

Cutting, J. E. (in preparation) Levels of processing in phonological fusion. Unpublished Ph.D. thesis, Yale University (Psychology).

Cutting, J. E. and R. S. Day. (1972) Fusion along an acoustic continuum. J. Acoust. Soc. Amer. 52, 175(A). (Also in Haskins Laboratories Status Report on Speech Research SR-28, 103-114.)

Day, R. S. (1968) Fusion in dichotic listening. Unpublished Ph.D. thesis, Stanford University (Psychology).

Lisker, L. (1957) Minimal cues separating /w, r, l, y/ in intervocalic position. Word 13, 257-267.

O'Connor, J. D., L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1957) Acoustic cues for the perception of initial /w, j, r, l/ in English. Word 13, 25-43.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to the left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Thorndike, E. L. and I. Lorge. (1944) The Teacher's Word Book of 30,000 Words. (New York: Teachers College Press).

A Speech Perception Paradox?: The Right-Ear Advantage and the Lag Effect

Robert A. Weeks[+]

For more than a century it has been held that language, at some level, is localized in one or the other cerebral hemisphere. In most people this dominant hemisphere is the left. For the past decade or so the use of dichotic speech perception experiments has enabled investigators to localize the unilateral nature of language at the level of speech perception, i.e., some level of the processing of speech signals is undertaken exclusively by the dominant hemisphere.

Two phenomena have been isolated in the dichotic speech situation. The first is the right-ear advantage, i.e., the right-ear input is correctly identified with a higher probability than the left-ear input. This advantage is maximum when stimulus presentations are simultaneous and declines as inputs are temporally staggered.

When inputs are staggered, another phenomenon emerges; this is the lag effect--the increased probability of identification of the lagging input relative to the probability of identification of the leading input. The lag effect is nonmonotonic, in that it increases as stimulus onset asynchronies are increased from zero to a maximum between 30 and 60 msec and decreases as stimulus onset asynchronies are increased further.

The theoretical interpretations of the two phenomena have evolved rather independently despite the similarities in the conditions necessary for the phenomena to occur. For example, both phenomena are manifested to the greatest degree if both inputs are linguistically deeply encoded, i.e., when phonemic information is acoustically carried via rapidly changing formant transitions such as stop-consonant vowel syllables (Studdert-Kennedy and Shankweiler, 1970; Studdert-Kennedy, Shankweiler, and Schulman, 1970) and nonsteady-state vowels in consonantal contexts (Weiss and House, 1973). It is in the theoretical interpretations of these phenomena that an apparent paradox emerges. The underlying processes inferred about one phenomenon preclude the underlying processes inferred

---

[+]The author is a student of Dr. M. T. Turvey, Haskins Laboratories and Department of Psychology, University of Connecticut. This paper is included because of its pertinence to current research at Haskins Laboratories.

Acknowledgment: Discussions with and critical comments from C. Darwin, A. Liberman, I. Mattingly, M. Studdert-Kennedy, and M. Turvey during earlier stages of this manuscript are gratefully acknowledged.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

in the interpretation of the other phenomenon. However, before pointing out the specifics of the apparent paradox, it is necessary to outline the interpretations of the right-ear advantage and the lag effect.

The right-ear advantage in dichotic speech perception is a well documented phenomenon (Kimura, 1961a, 1961b, 1967; Shankweiler and Studdert-Kennedy, 1966; Curry, 1967; Curry and Rutherford, 1967; Kimura and Folb, 1968; Darwin, 1969; Studdert-Kennedy and Shankweiler, 1970), and efforts to converge upon an understanding of the locus and the nature of the right-ear advantage have been relatively successful (Kimura, 1961b, 1964; Milner, Taylor, and Sperry, 1968; Kirstein and Shankweiler, 1969; Studdert-Kennedy and Shankweiler, 1970). In general, these efforts have led to the view of a unilateral linguistic processor in the left hemisphere that receives acoustic inputs from both ears. Thus, under dichotic conditions, during the course of transmission to the linguistic processor the signal from the ipsilateral (left) ear undergoes a relatively greater loss than the signal from the contralateral (right) ear because it must first travel to the right hemisphere before it is transmitted to the left hemisphere (Kimura, 1961b; Studdert-Kennedy and Shankweiler, 1970; Berlin, Lowe-Bell, Cullen, Thompson, and Loovis, in press-a). The ipsilateral pathways have been deemphasized in these interpretations as it is assumed that there is a functional prepotency of the contralateral pathway from the right ear to the left hemisphere (Kimura, 1961b). This assumption has drawn support from physiological evidence that the contribution of contralateral auditory pathways is greater than that of ipsilateral pathways (Bocca, Calearo, Cassinari, and Migliavacca, 1955) and evidence that the ipsilateral signal is inhibited during dichotic stimulation (Milner et al., 1968). Thus, the present view has localized the origin of the right-ear advantage immediately before, within, or immediately after the interface between auditory processing and initial linguistic processing.

On the foregoing understanding of the locus of the right-ear advantage, two explanatory hypotheses can be considered. The first hypothesis invokes the concept of queueing, and there are two versions of this hypothesis. The first version localizes the delay within the linguistic processor: the processing of the left-ear input is detained at some point during the processing of the left-ear input, and during this detention some ipsilateral information is lost.

A queueing hypothesis of this kind is untenable on several grounds. First, if input selection is based upon transmission line--the direct contralateral or the indirect interhemispheric line--it would be necessary, at the point of selection, for channel information to be available, and used, by the mechanisms responsible for selection. Contrary to this notion is the finding that "blend" errors, which indicate loss of local sign, occur at a rate greater than chance in dichotic speech experiments (Studdert-Kennedy and Shankweiler, 1970). Second, if selection is based upon time of arrival to the linguistic processor, it would become necessary to invoke the idea of a processing bias for the initial member of a pair of linguistic events. On our present understanding of central perceptual processes this seems unlikely; centrally, the advantage accrues to the later arriving event (Turvey, 1973) (I will return to this later). Third, and most devastating to the viability of this kind of queueing hypothesis, is the finding that the right-ear advantage can be attenuated or reversed by manipulating the intensity of the respective inputs (Thompson, Stafford, Cullen, Hughes, Lowe-Bell, and Berlin, 1972; Brady-Wood and Shankweiler, 1973). If the right-ear advantage originates within the linguistic processor, it would be expected that only

manipulations of acoustic dimensions which carry phonemic information could affect linguistic processing. Dichotic backward masking (to be considered shortly) is presumed to be of central origin, and available evidence suggests that intensity has little, if any, effect upon the temporal course or amount of backward masking (Pisoni, 1972). Thus it appears that the signal transmitted interhemispherically is acoustically encoded, and further, that the right-ear advantage originates immediately before or within, but not after, the interface between the auditory signal and initial linguistic processing.

A second version of the queueing hypothesis ascribes the right-ear advantage to a loss suffered in the additional time necessary for the left-ear input to reach the dominant hemisphere. Implicit in this view, of course, is the assumption that the left-ear input must tranverse a longer distance. This variety of the queueing hypothesis is essentially identical to the second general hypothesis, the distance hypothesis. The distance hypothesis, simply stated, is that the right-ear advantage is due to a loss in signal clarity suffered in the additional distance traveled by the left-ear input as it is shuttled to the dominant hemisphere. Implicit in this view is the assumption that the left-ear input must consume more time than the right-ear input during its journey to the speech processor.

The difference between these hypotheses is merely one of emphasis on the origin of the impairment of the left-ear input; both ascribe the loss and resultant right-ear advantage to stresses incurred due to structural properties of the interhemispheric transmission system. These two hypotheses, therefore, can be regarded as one which we may call the "degradation hypothesis." Quite simply, the degradation hypothesis views the right-ear advantage as the result of the poorer input that the ipsilateral signal presents to the higher-level feature extractors in the dominant hemisphere. Although unlabeled previously, the degradation hypothesis appears to be currently favored (Studdert-Kennedy and Shankweiler, 1970; Berlin et al., in press-a; in press-b).

The degradation hypothesis attributes the right-ear advantage to an impairment in the left-ear signal incurred during transmission, and although it attributes that loss to the additional time and/or distance necessary for that transmission, it is the impairment which "explains" right-ear advantage. But the implied late arrival to the dominant hemisphere of the ipsilateral input should not be overlooked. If the degradation hypothesis is to be considered, it should be noted that in the context of the hypothesis, the ipsilateral input does arrive at the dominant hemisphere at some finite time, say, α msec, later than the contralateral input. I shall return to this point shortly, but let us first consider the lag effect.

The lag effect, like the right-ear advantage, is a well documented phenomenon of dichotic speech perception (Berlin et al., 1970, in press-a; Kirstein, 1970, 1971; Lowe, Cullen, Thompson, Berlin, Kirkpatrick, and Ryan, 1970; Studdert-Kennedy et al., 1970; Darwin, 1971; Porter, 1971). The occurrence of backward masking with speech stimuli has provoked comparisons between visual and speech masking (Studdert-Kennedy et al., 1970; Darwin, 1971). In vision, dichoptic masking is asymmetrical, with significantly more backward than forward masking and with contour interaction a necessary condition (Kahneman, 1968; Turvey, 1973). Probably a distinction between central and peripheral masking is more valid than the forward-backward distinction (Turvey, 1973). The favored interpretation of central masking is that it is fundamentally backward in effect and is due to the

interruption or distortion of central perceptual processing (Kahneman, 1968; Spencer and Shuntich, 1970). Central backward masking is viewed as the interruption of ongoing feature-extraction processes by a more recent informational event, and this type of masking phenomenon follows a time course where first-event processing appears to be maximally susceptible to disruption at some non-zero stimulus onset asynchrony (Weisstein, 1969; Turvey, 1973). An analogous interpretation has been proposed as an explanation of the dichotic lag effect (Studdert-Kennedy et al., 1970; Darwin, 1971; Porter, 1971; Berlin et al., in press-a).

In the context of the foregoing interpretations of the right-ear advantage and the lag effect the apparent paradox emerges. Although at present the degradation hypothesis of right-ear advantage seems more in harmony with available data, both the degradation and the queueing hypothesis assume that the ipsilateral input arrives at the linguistic processing dominant hemisphere some finite time later than the contralateral input. Given this assumption, the conditions are present for central backward masking to occur, yet it is held that the right-ear advantage and the lag effect are independent (Kirstein, 1970; Berlin et al., in press-a).

In the queueing hypothesis, the extraction operations by the left hemisphere upon the ipsilateral input are delayed at least at one stage during the course of contralateral input processing. However, in the interruption interpretation of the lag effect, the left-ear input should interrupt the processing of the right-ear input. Thus, the view of the interruption of leading-stimulus processing as an interpretation of the lag effect is in direct conflict with the view of detention of leading-stimulus processing as an interpretation of the right-ear advantage.

Therefore, when the evidence contrary to the queueing hypothesis as an interpretation of the right-ear advantage and the evidence in harmony with the interruption hypothesis as an interpretation of the lag effect are considered, the queueing hypothesis clearly must be rejected.

The apparent paradox still exists, however, as the degradation hypothesis places the paradox in a more subtle context. Although the mechanisms underlying the right-ear advantage and the lag effect are assumed to differ and be independent (Kirstein, 1970; Studdert-Kennedy et al., 1970; Berlin et al., in press-a), the implication, from the context of the right-ear advantage interpretations, remains that for dichotic simultaneous presentations the ipsilateral input arrives at some finite time, α, later at the dominant hemisphere. On this implication, the temporal course of the lag effect should differ with respect to which ear is lagging at presentation. For example, if the lag effect attains a maximum with a stimulus onset asynchrony of 40 msec with respect to arrival at the speech processor, the maxima with respect to presentation should be manifested at $(40-\alpha)$ msec when the left-ear input is lagging and at $(40+\alpha)$ msec when the right-ear input is lagging. Yet available evidence suggests that the lag effect maxima are symmetrical about the simultaneous presentation situation (Kirstein, 1971; Berlin et al., in press-a).

It is inferred from the views of the right-ear advantage that simultaneity of inputs at presentation is not retained upon arrival to the left hemisphere, and yet the symmetry of the lag effect suggests that the speech processing system

is operating around the simultaneous presentation situation as if the inputs arrived at the central speech processor simultaneously. Thus, there appears to be no graceful exit from the apparent paradox which interpretations of the right-ear advantage and the lag effect have generated.

Are we to hypothesize that the time, α, is so small that it is insignificant? This seems unlikely on at least two grounds. First, it is probable that the left-ear input is encoded in some fashion before interhemispheric transfer and recoded upon arrival at the left hemisphere. Little, if any, encoding processes can occur prior to arrival at the dominant hemisphere if α is to be small. Second, estimates of interhemispheric transfer--between 10 and 35 msec (Bremer, 1958; Grafstein, 1959; Teitelbaum, Sharples, and Byck, 1968; Moscovitch and Catlin, 1970)--indicate that interhemispheric transfer time, although possibly small, is not insignificant.

An alternative resolution to the paradox is to hypothesize a functional simultaneity point at the speech processor where the system adjusts to the time difference, α, and where simultaneously presented inputs are treated as if those inputs arrived at the dominant hemisphere simultaneously. This resolution is similar to an interpretation proposed by Berlin et al. (in press-a). Although understating the apparent paradox, they are, until the present, the sole investigators to address the paradox. Assuming that interhemispheric transmission time is less than 10-15 msec, they hypothesize that if two stimuli arrive at the linguistic processor within 10-15 msec of each other, the inputs are treated as if they arrived simultaneously. However, Berlin et al. do not go far enough to resolve the apparent paradox since their concept of a "psychological moment" is bidirectional. If it is assumed that interhemispheric transfer time is α, that backward masking does not occur if inputs arrive at the linguistic processor within β msec of each other, and that $\beta > \alpha$, the maxima of the lag effect should still remain asymmetrical about the simultaneous presentation condition. Although they can "explain" why backward masking should not occur in the simultaneous presentation situation, they leave unexplained the temporal course of the lag effect. For example, if the lag effect attains a maximum with a stimulus onset asynchrony of 40 msec with respect to arrival at the speech processor, the maxima with respect to presentation should occur at $(40-\alpha)$ msec when the left-ear input is lagging and at $(40+\alpha)$ msec when the right-ear input is lagging. In addition, no lag effect should be manifested between stimulus onset asynchronies of $(\beta+\alpha)$ msec when the right-ear input is lagging at presentation and $(\beta-\alpha)$ msec when the left-ear input is lagging at presentation. According to available evidence the temporal course of the lag effect does not follow the foregoing predictions (Kirstein, 1971; Berlin et al., in press-a).

Thus, a hypothesized "psychological moment," such as suggested by Berlin, must possess a unilateral character; the left hemisphere must adjust for only the late arrival of left-ear input. Berlin and colleagues do not take this hypothetical step, possibly because such a conceptualization would bestow an intuitively unlikely power upon the speech processor. Given that the lag effect is congruent with what is accepted about central processing in general, it seems most probable that the interpretations of the right-ear advantage are deficient and in need of reevaluation. Perhaps, the role of the ipsilateral pathways has been dismissed too readily. In any event, it appears that a comfortable resolution to the apparent paradox is not presently available.

REFERENCES

Berlin, C. I., S. S. Lowe-Bell, J. K. Cullen, C. L. Thompson, and C. F. Loovis. (in press-a) Dichotic speech perception: an interpretation of right-ear advantage and temporal offset effect. J. Acoust. Soc. Amer.

Berlin, C. I., R. J. Porter, S. S. Lowe-Bell, H. L. Berlin, C. L. Thompson, and L. F. Hughes. (in press-b) Dichotic signs of the recognition of speech elements in normals, temporal lobectomees, and hemispherectomees. IEEE Trans.

Berlin, C. I., M. E. Willet, C. L. Thompson, J. K. Cullen, and S. S. Lowe. (1970) Voiceless versus voiced CV perception in dichotic and monotic listening. J. Acoust. Soc. Amer. 47, 75(A).

Bocca, E., C. Calearo, V. Cassinari, and F. Migliavacca. (1955) Testing "cortical" hearing in temporal lobe tumors. Acta Oto-Laryngologica 45, 289-304.

Brady-Wood, S. and D. Shankweiler. (1973) Effects of attenuation of one of two channels on perception of opposing pairs of nonsense syllables when monotically and dichotically presented. Paper presented at the 85th meeting of the Acoustical Society of America, Boston, Mass.

Bremer, F. (1958) Physiology of the corpus collosum. Research Publications of the Association for Nervous and Mental Diseases 36, 424-448.

Curry, F. K. W. (1967) A comparison of left-handed and right-handed subjects on verbal and nonverbal dichotic listening tasks. Cortex 3, 343-352.

Curry, F. K. W. and D. R. Rutherford. (1967) Recognition and recall of dichotically presented verbal stimuli by right- and left-handed persons. Neuropsychologica 5, 119-126.

Darwin, C. J. (1969) Laterality effects in the recall of steady-state and transient speech sounds. J. Acoust. Soc. Amer. 46, 114(A).

Darwin, C. J. (1971) Dichotic backward masking of complex sounds. Quart. J. Exp. Psychol. 23, 386-392.

Grafstein, D. (1959) Organization of callosal connections in supra-sylvian gyrus of cat. J. Neurophysiol. 22, 504-515.

Kahneman, D. (1968) Method, findings, and theory in studies of visual masking. Psychol. Bull. 70, 404-425.

Kimura, D. (1961-a) Some effects of temporal lobe damage on auditory perception. Canad. J. Psychol. 51, 156-165.

Kimura, D. (1961-b) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1964) Left-right differences in the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.

Kimura, D. (1967) Functional asymmetry of the brain in dichotic listening. Cortex 3, 163-178.

Kimura, D. and S. Folb. (1968) Neural processing of backward speech sounds. Science 161, 395-396.

Kirstein, E. F. (1970) Selective listening for temporally staggered dichotic CV syllables. J. Acoust. Amer. 48, 95(A).

Kirstein, E. F. (1971) Temporal factors in perception of dichotically presented stop consonants and vowels. Ph.D. dissertation, University of Connecticut.

Kirstein, E. F. and D. Shankweiler. (1969) Selective listening for dichotically presented consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-17/18, 133-141.

Lowe, S. S., J. K. Cullen, C. L. Thompson, C. I. Berlin, L. L. Kirkpatrick, and J. T. Ryan. (1970) Dichotic and monotic simultaneous and time-staggered speech. J. Acoust. Soc. Amer. 47, 76(A).

Milner, B., L. Taylor, and R. W. Sperry. (1968) Lateralized suppression of dichotically-presented digits after commissural section in man. Science 161, 184-185.

Moscovitch, M. and J. Catlin. (1970) Interhemispheric transmission of information: measurement in normal man. Psychon. Sci. 18, 211-213.

Pisoni, D. B. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-31/32, 83-92.

Porter, R. J. (1971) The effect of temporal overlap on the perception of dichotically and monotically presented CV syllables. Paper presented at the 81st meeting of the Acoustical Society of America, Washington, D. C.

Shankweiler, D. and M. Studdert-Kennedy. (1966) Lateral differences in perception of dichotically presented synthetic consonant-vowel syllables and steady-state vowels. J. Acoust. Soc. Amer. 39, 1256(A).

Spencer, T. J. and R. Shuntich. (1970) Evidence for an interruption theory of backward masking. J. Exp. Psychol. 85, 198-203.

Studdert-Kennedy, M. and D. Shankweiler. (1970) Hemispheric specialization for speech perception. J. Acoust. Soc. Amer. 48, 579-594.

Studdert-Kennedy, M., D. Shankweiler, and S. Schulman. (1970) Opposed effects of a delayed channel on perception of dichotically and monotically presented CV syllables. J. Acoust. Soc. Amer. 48, 599-692.

Teitelbaum, H., S. K. Sharples, and R. Byck. (1968) Role of somatosensory cortex in interhemispheric transfer of tactile habits. J. Comp. Physiol. Psychol. 66, 623-632.

Thompson, C. L., M. R. Stafford, J. K. Cullen, L. F. Hughes, S. S. Lowe-Bell, and C. I. Berlin. (1972) Interaural intensity differences in dichotic speech perception. Paper presented at the 83rd meeting of the Acoustical Society of America, Buffalo, N. Y.

Turvey, M. T. (1973) On peripheral and central processes in vision: inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Weiss, M. S. and A. S. House. (1973) Perception of dichotically presented vowels. J. Acoust. Soc. Amer. 53, 51-58.

Weisstein, N. A. (1969) What the frog's eye tells the human brain: single cell analyses in the human visual system. Psychol. Bull. 72, 157-176.

Perception of Speech and Nonspeech, with and without Transitions

James E. Cutting[+]
Haskins Laboratories, New Haven, Conn.

What role do rapidly changing energy patterns play in the perception of auditory signals? The work of Pollack (1968) and Nabelek and Hirsh (1968) have revealed much about how these signals are perceived in isolation, but we still remain relatively uninformed as to their role in more complex signals. We know that speech sounds, which typically have a wealth of moving formants, are processed differently than nonspeech signals, which tend not to have such variation. The best evidence for this differential processing comes from dichotic listening tasks. Dichotic tasks involving speech stimuli generally yield right-ear advantages, while dichotic tasks using nonspeech stimuli generally yield left-ear advantages. These results are usually interpreted in terms of processing capabilities in the two hemispheres of the brain (see Kimura, 1967; Semmes, 1968; and Day, in press).

Differential processing can also occur within the category of speech. For example, speech sounds with transitions are processed differently than those without transitions (Shankweiler and Studdert-Kennedy, 1967; Day and Vigorito, 1973). Are nonspeech sounds with and without transitions also processed differently? The data of Halperin, Nachshon, and Carmon (1973) suggest so, but is the differential processing for stimuli with and without transitions of the same nature for both speech and nonspeech? Perhaps the processing of rapid frequency transitions is done independently of the class of stimuli.

A dichotic listening task was devised to test the role of formant transitions in speech and nonspeech signals. A temporal order judgment task was selected which did not require the subjects to give verbal labels to the stimuli. Instead, subjects merely recognized the item which began first in a dichotic pair.

METHOD

Stimuli. Nine speech stimuli and nine nonspeech stimuli were prepared using the facilities of the Haskins Laboratories. Speech stimuli were prepared on the parallel resonance synthesizer and included three steady-state vowel (V) stimuli [i, ae, ɔ] and six consonant-vowel (CV) stimuli [bi, gi, bae, gae, bɔ, gɔ]. Each of the speech stimuli was made to have the same falling pitch contour. They were then transferred to the pulse code modulation (PCM) system for the

---

[+]Also Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

preparation of dichotic tapes (Cooper and Mattingly, 1969). Nonspeech stimuli were constructed out of pure tones and were assembled on the PCM system. Each consisted of three tones superimposed on top of one another. The frequency of the tones in each of the nonspeech stimuli corresponded to the central frequency of the formants in the speech stimuli. Thus, one nonspeech stimulus was composed of tones of 660, 1720, and 2410 Hz, the same median values for the formants of the vowel [ae]. Three of the nonspeech stimuli consisted of steady-state pure tones whose values corresponded to the central formant frequencies of the vowels [i, ae, ɔ]. These three nonspeech vowel-like stimuli did not sound like speech and will be designated NS (v). Six other nonspeech stimuli had portions identical to the steady-state tone stimuli [NS (v)] but they had frequency modulations for the first 50 msec. The modulations rose or fell to the steady-state frequency of the tones, and corresponded to transitions in the CV stimuli. Since they resembled the CV stimuli in their general structure but did not sound like speech they will be designated NS (cv). Amplitudes of the component tones in the nonspeech stimuli were adjusted to match the relative amplitudes of the formants in the speech stimuli. Speech and nonspeech stimuli were identical in intensity and duration. Figure 1 shows schematic representations of four stimuli used in the present study, one from each class. Thick bars indicate formants in the speech stimuli while narrow lines indicate the pure tones in the nonspeech stimuli. The speech stimuli shown are [bae] and [ae]. The nonspeech stimuli are those which correspond to them.

Tapes and procedure. Trials for the dichotic temporal order judgment task were constructed so that subjects would not have to identify stimuli on a particular trial. Instead, they would recognize the stimulus in a given dichotic pair by means of a probe stimulus. A trial, therefore, consisted of a dichotic pair with a temporal onset asynchrony of 50 msec, followed by one second of silence, followed by a diotic stimulus which was one of the members of the dichotic pair. Subjects were instructed to regard the diotic stimulus as a probe which asked the question: "Is this the stimulus which began first?" Figure 2 shows a schematic representation of two such trials. Consider sample trial 1. Stimulus 1 begins before Stimulus 2 by 50 msec, and the probe stimulus is Stimulus 1. Since the probe is the stimulus which began first, the correct response is yes. In sample trial 2, the dichotic pair is the same as in trial 1, but the probe stimulus is different. Since Stimulus 2 did not begin before Stimulus 1, the correct response for trial 2 is no.

Four tapes were constructed, one for each class of stimuli. Each tape consisted of 48 trials. The stimuli used in each trial were always selected from the same class of stimuli. CV trials were constructed out of CV stimuli which shared neither the same vowel nor the same consonant: thus, for example, [bi] was paired with [gae] and [gɔ]. In V trials, different vowel stimuli were paired: for example, [i] was paired with [ae] and [ɔ]. NS (cv) trials were constructed using the same rules applied to CV stimuli: the two stimuli in the dichotic pair could share neither the same tone structure nor have pitch modulations which correspond to the transition of the same stop consonant. The rules for the construction of NS (v) pairs were the same as those for V trials. Of course, stimuli in the dichotic pair were counterbalanced for leading and lagging position. The probe stimulus chosen for each trial and the channel assignments of the stimuli in the dichotic pair were also counterbalanced in a random sequence of trials.

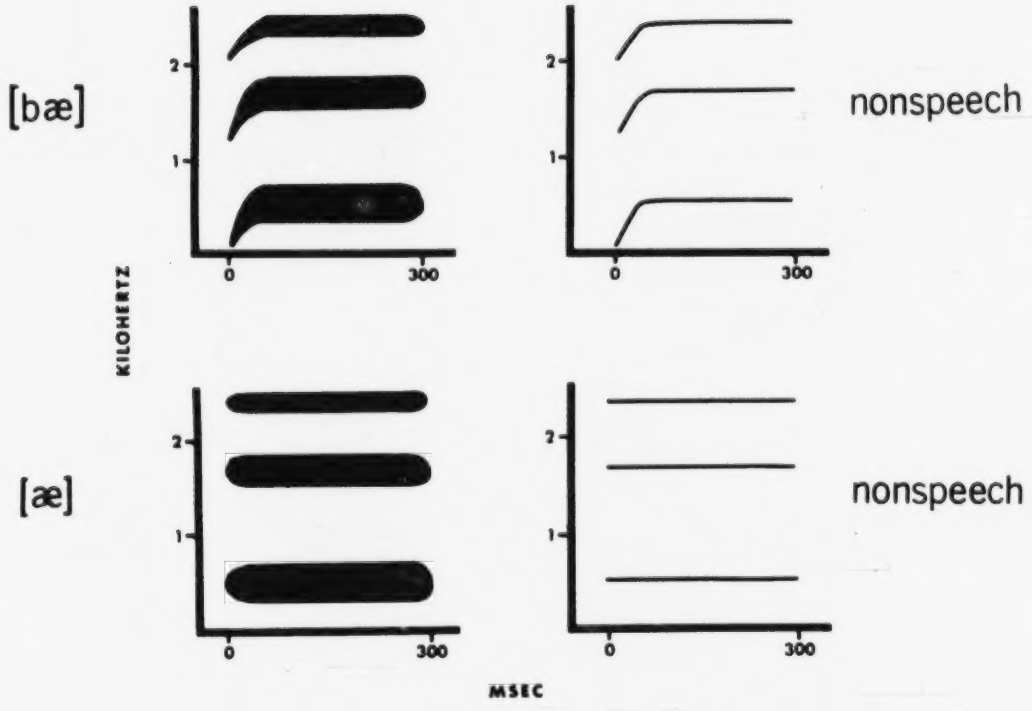


Figure 1: Schematic spectrograms of sample speech and nonspeech stimuli, with and without transitions.

Figure 2

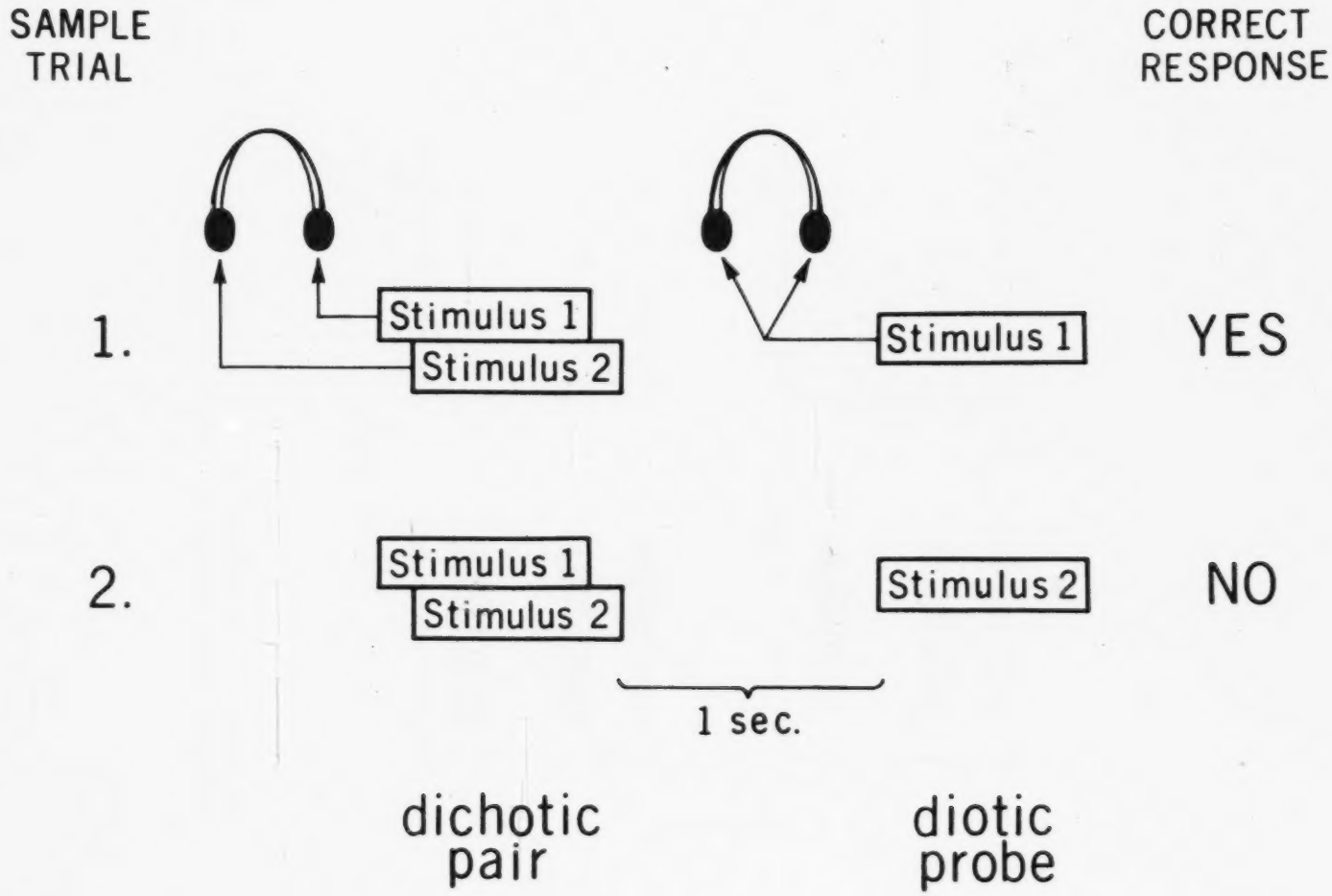


Figure 2: The paradigm used for temporal order judgments.

Subjects and apparatus. Sixteen Yale University undergraduates listened to all four tapes. They were all right-handed native American English speakers with no history of hearing difficulty. They were run in four groups of four, listening to tapes played on an Ampex AG500 dual track tape recorder sent through a listening station to Grason Stadler earphones (model TDH39-300Z). Subjects listened to each tape twice, reversing the earphones after one pass through the tape. The order of channel assignments was counterbalanced across subjects. Each group listened to the four tapes in a different order, determined by a balanced Latin Square design. Subjects listened to a total of 384 trials consisting of a dichotic pair and a diotic probe. The response for each trial was yes or no.

RESULTS

In general the task was quite difficult: overall performance for all trials and all types of stimuli was 65 percent correct. Overall performance for each of the four types of stimuli was comparable. The average score for each was between 64 and 66 percent, with no significant differences among them.

The pattern of ear advantages is quite interesting, but before discussing them, it is necessary to note how the results were scored. Consider again the sample trials in Figure 2. The correct response for sample trial 1 is yes, while the correct response for trial 2 is no. If in the dichotic pair Stimulus 1 was presented to the right ear and Stimulus 2 to the left ear, and if the subject responded yes for the first trial and no for the second, he would have both correct. This would have been scored as two correct responses for the right-ear leading stimulus. If the subject, on the other hand, had responded no and yes for the sample trials respectively, both would be wrong and his score for the right-ear leading stimulus would be docked for two incorrect responses. [Of course if in the dichotic pair the channels had been reversed and Stimulus 1 was presented to the left ear and Stimulus 2 to the right, the logic would be entirely reversed.] Now let us consider each stimulus class in turn.

CV trials. There was a large significant ear difference for the consonant-vowel trials. When the leading stimulus of the dichotic pair was presented to the right ear, subjects were 72 percent correct in responding to the probe stimulus. When the leading stimulus was presented to the left ear, on the other hand, subjects were only 57 percent correct, yielding a net 15 percent right-ear advantage. All subjects showed results in this direction ($z = 3.75$, $p < .0001$).

V trials. There was a significant ear difference for the steady-state vowels as well. When the leading stimulus was presented to the right ear, subjects were 69 percent correct, while they were only 62 percent correct when the leading stimulus was presented to the left, yielding a net 7 percent right-ear advantage. Thirteen of 16 subjects showed results in this direction ($z = 2.25$, $p < .02$).

NS (cv) trials. No significant ear difference was found for the nonspeech stimuli most resembling the CV stimuli. Ear scores were 65 and 64 percent correct for trials in which the right and left ear stimuli led, yielding a net 1 percent right-ear advantage. In a post-experiment interview no subject reported that the NS (cv) stimuli sounded like speech.

NS (v) trials. A small left-ear advantage was found for the nonspeech stimuli which were most like vowels. When the steady-state tone stimulus led in the right ear subjects were 63 percent correct, while they were 68 percent correct when it led in the left-ear, yielding a net 5 percent left-ear advantage. The ear advantage for these trials was not significant. Again, no subject reported that the NS (v) stimuli sounded like speech.

The results of each of the four conditions is shown in Figure 3. Observe the relationships of the ear difference scores for the different classes of stimuli.

Speech vs. nonspeech. The dimension of speech versus nonspeech proved to be a significant factor in the temporal order judgment of these stimuli. Subjects showed a large right-ear advantage for the speech stimuli. Averaging the ear scores for the CV and V stimuli we find that there was an 11 percent right-ear advantage. This superiority of the right-ear leading stimuli was highly significant (F (1,15) = 19.3, $p<.001$). Nonspeech stimuli, however, showed a different pattern of results. Collapsing over NS (cv) and NS (v) stimuli subjects had a 2 percent advantage for the left-ear leading stimulus. This small effect was not significant itself, but the difference in the two classes of stimuli, speech versus nonspeech, was highly significant (F (1,15) = 19.4, $p<.001$).

Stimuli with and without transitions. Independent of the dimension of speech versus nonspeech, transitions played a definite role in temporal order judgments. If we collapse across the CV and NS (cv) stimuli we find that subjects had a marked difference in their ear scores. The net ear difference for these trials was 8 percent in favor of the right-ear leading stimuli. Collapsing over the other two sets of stimuli we do not find this ear difference. The V and NS (v) stimuli yielded a net difference of only 1 percent in favor of the right-ear leading stimulus. The difference in the results for the stimuli with transitions and those without transitions was significant (F (1,15) = 9.45, $p<.01$). No other higher order interactions were significant: for example, there was no difference in the relationships of ear scores for the CV and V stimuli as opposed to the NS (cv) and NS (v) stimuli.

DISCUSSION

Differences in the perception of speech and nonspeech have been well documented. Perhaps the best source of information about these two types of stimuli comes from dichotic listening tasks. In general, speech tasks, whether identification or temporal order judgment, yield right-ear advantages (Kimura, 1961; Day and Cutting, 1971). Nonspeech tasks, on the other hand, whether identification or temporal order judgment, generally yield left-ear advantages (Chaney and Webster, 1966; Day and Cutting, 1971). The results of the present study are in accord with these findings. Speech stimuli [CV and V] yielded a highly significant advantage to stimuli leading in the right ear, nonspeech stimuli [NS (cv) and NS (v)] yielded a small advantage to stimuli leading in the left ear, and while this ear advantage for nonspeech stimuli was not significant, it was significantly different from the speech stimuli. Thus, as in previous studies we can assume that the processing was different for the two types of stimuli, each type of stimuli requiring different general amounts of processing in each hemisphere. Since speech/nonspeech differences have been given so much attention in the past, let us turn to the other, and perhaps more interesting general result.

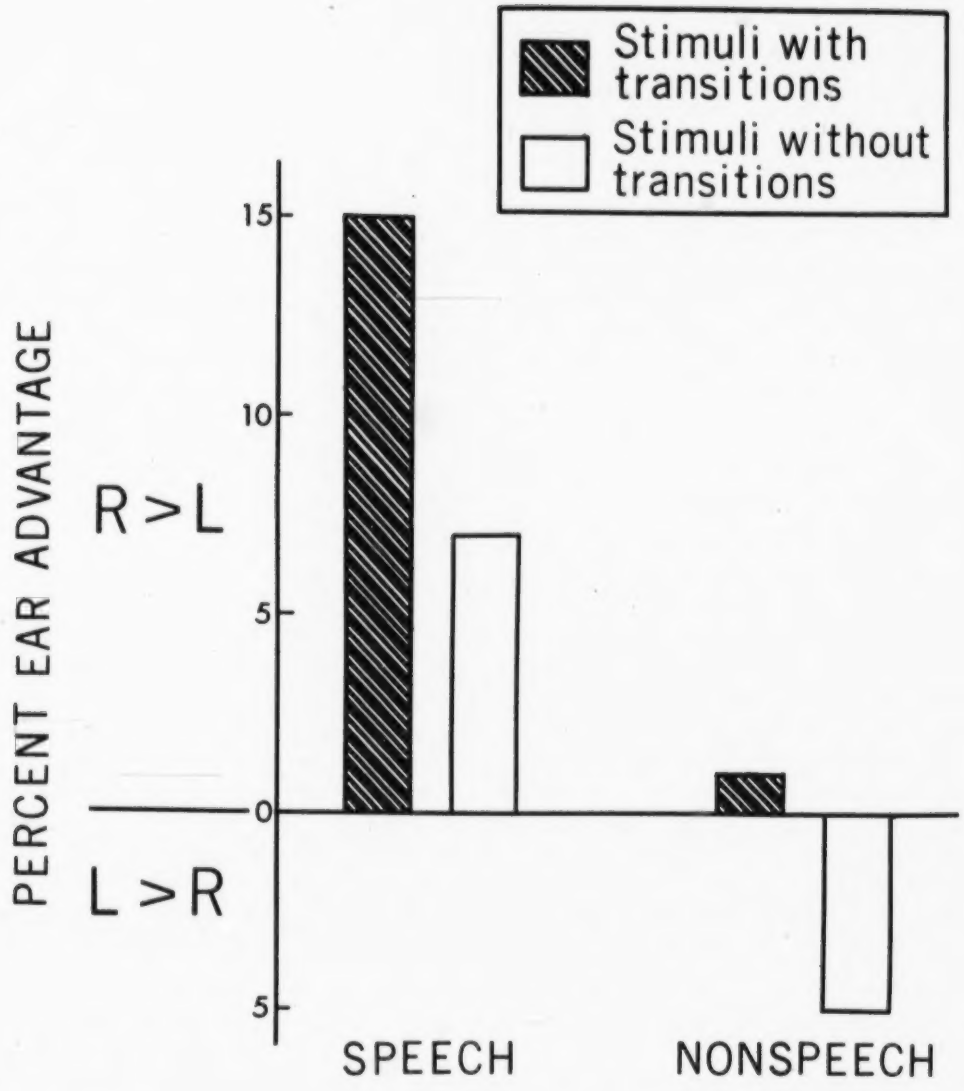


Figure 3: Ear advantages for speech and nonspeech stimuli, with and without transitions.

Stimuli with transitions were perceived differently than stimuli without transitions regardless of whether or not they were speech stimuli or nonspeech stimuli. It appears that a stimulus which contains rapid changes in frequency requires special processing in the left hemisphere, and that this processing is independent of whether the stimulus is classified as speech or not. How might this be explained?

The processing capabilities of the left hemisphere appear to be fundamentally different from those of the right hemisphere. Most of the differences appear to be related to language, and range through all levels of language from phonetics to semantics. Perhaps there is a specialization of the left hemisphere system for certain purely auditory processes. One such perceptual process may be the analysis of rapidly changing frequency modulations, independent of whether they are formant transitions in speech stimuli or rapid pitch sweeps in nonspeech stimuli. If speech and language processing is primarily confined to one hemishpere, it would be advantageous to have certain subsidiary systems in that hemisphere to assist in the demodulation of the incoming speech signal. One such subsystem might be an analyzer which tracks rapidly changing frequencies such as those found in speech. Such a subsystem would not be needed in a hemisphere geared for nonlinguistic analysis since very few nonlinguistic signals have rapidly changing frequencies. The notion of a transition analyzer in the left hemisphere is in keeping with Semmes' (1968) views of hemisphere differences in processing. She states that left hemisphere function is characterized by "focal" organization, whereas the right hemisphere is characterized by a more "diffuse" organization. Certainly the analysis of rapid pitch modulations is a "focal" task requiring very discrete detectors. It is not surprising that we should find such an analysis performed by the left hemisphere system.

In the present study we have two variables: whether or not the stimulus is speech, and whether or not the stimulus has transitions. We know that speech processing is primarily a left-hemisphere task, and suspect that the analysis of transitions is also a left-hemisphere task. The two processes appear to be independent of one another, and if they are independent they are also additive. CV stimuli, both [+ speech] and [+ transition], yield a large right-ear/left-hemisphere advantage since the two variables favor left-hemisphere processing. V and NS (cv) stimuli, however, have only one positive value on the two dimensions and thus yield smaller ear difference scores: V stimuli are [+ speech] and [- transition] and NS (cv) stimuli are [- speech] and [+ transition]. NS (v) stimuli are both [- speech] and [- transition], and consequently yield a left-ear advantage. Note that speech/nonspeech is a more potent dimension than is transition/nontransition: V stimuli yield a larger right-ear advantage than do NS (cv) stimuli.

The results of the present study are similar to those of Darwin (1971). He found that fricatives with formant transitions yielded a right-ear advantage, while the same fricatives without transitions yielded no advantage in either direction. Perhaps Darwin's result may be explained by the same transition analyzer which I have postulated here. It seems reasonable to suppose that transitional energy is analyzed independently of whether the stimulus is speech or nonspeech. Otherwise, the system must necessarily make an early decision to determine whether or not a stimulus is speech, before starting to analyze its transitions. Such a process would be unnecessarily cumbersome, if not untenable.

Perhaps this transition analyzer in the left hemisphere can be invoked to a greater or lesser degree according to the stimuli being processed; or perhaps it shows a differential advantage over the processors of the right-hemisphere system according to how rapid the transitions are. In either case it would account for the results of Cutting (1972) and Day and Vigorito (1972). Both studies found a continuum of ear advantages with stop consonants yielding a large right-ear advantage, liquids yielding a reduced right-ear advantage, and vowels yielding either no ear advantage or a left-ear advantage. Transitions in stop consonants are quite rapid, on the order of about 50 msec. Liquid transitions, however, are slightly longer, usually about 100 msec. Steady-state vowels, of course, do not have transitions. Perhaps in the case of the liquids, the transition analyzer in the left hemisphere has only a moderate advantage over the processors in the right hemisphere, whereas the advantage is much more marked in the case of stop consonants. Thus, the differences between the ear scores for stops, liquids, vowels, and other phoneme classes may be a function of the differential involvement of a transition analyzer in the left hemisphere.

SUMMARY AND CONCLUSION

Subjects judged the temporal order of similar speech and nonspeech stimuli, with and without transitions, in a task which did not require them to label the stimuli. Results showed a large right-ear advantage for judging temporal order of speech stimuli, and a small but significantly different left-ear advantage for the judgment of nonspeech stimuli. In addition to, and independent of, the dimension of speech/nonspeech, there was a significant difference in the ear scores for stimuli with and without transitions: stimuli with formant transitions yielded a significant right-ear advantage, while those without transitions yielded no ear advantage in either direction. These results may be interpreted as an indication of a special nonlinguistic processor in the left hemisphere. This special processor may be thought of as a subsystem to the language processor whose function is to analyze rapid frequency modulations in all auditory signals, whether they are speech or not.

REFERENCES

Chaney, R. B. and J. C. Webster. (1966) Information in certain multidimensional sounds. J. Acoust. Soc. Amer. 40, 447-455.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cutting, J. E. (1972) A parallel between degree of encodedness and the ear advantage: evidence from an ear-monitoring task. J. Acoust. Soc. Amer. 53, 358(A). (Also in Haskins Laboratories Status Report on Speech Research SR-29/30, 61-68 as: A parallel between encodedness and the magnitude of the right-ear advantage.)

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exper. Psychol. 23, 46-62.

Day, R. S. (in press) Engaging and disengaging the speech processor. In Hemispheric Asymmetry of Function, ed. by Marcel Kinsbourne. (London: Tavistock).

Day, R. S. and J. E. Cutting. (1971) What constitutes perceptual competition in dichotic listening? Paper presented at the Eastern Psychological Association meeting, New York, April.

Day, R. S. and J. M. Vigorito. (1972) A parallel between encodedness and the ear advantage: evidence from a temporal order judgment task. J. Acoust. Soc. Amer. 53, 358(A). (Also in Haskins Laboratories Status Report on Speech Research SR-31/32.)

Halperin, Y., I. Nachshon, and A. Carmon. (1973) Shift in ear superiority in dichotic listening to temporally patterned nonverbal stimuli. J. Acoust. Soc. Amer. 53, 46-50.

Kimura, D. (1961) Cerebral dominance and the perception of verbal stimuli. Canad. J. Psychol. 15, 166-171.

Kimura, D. (1967) Dual functional asymmetry of brain in visual perception. Neuropsychol. 4, 275-285.

Nabelek, I. and I. J. Hirsh. (1968) On the discrimination of frequency transitions. J. Acoust. Soc. Amer. 45, 1510-1519.

Pollack, I. (1968) Detection of rate of change of auditory frequency. J. Exper. Psychol. 77, 535-541.

Semmes, J. (1968) Hemispheric specialization: a possible clue to mechanism. Neuropsychol. 5, 11-26.

Shankweiler, D. P. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exper. Psychol. 19, 59-63.

Dichotic Release from Masking for Speech

Timothy C. Rand[+]
Haskins Laboratories, New Haven, Conn.

Several investigators have observed that presenting the first formant (F1) of synthetic speech to one ear and the second formant (F2) to the other ear results in fusion (Broadbent, 1955; Broadbent and Ladefoged, 1957; Halwes, 1969). Two experiments are reported here that take advantage of this dichotic phenomenon in exploring the perceptual interaction between different formants and formant transitions at varying amplitudes.

It is well known in auditory psychophysics that a low-frequency tone is an effective masker of a higher-frequency tone. Relatively little has been done along these lines with speech, but one might suspect, along with Flanagan and Saslow (1958), that an analogous statement could be made for speech stimuli. Whereas strong low-frequency speech components may mask higher-frequency components under binaural listening conditions, the present results indicate that dichotic presentation produces a release from this masking.

EXPERIMENT I

In the first of the two experiments, perception of synthetic speech syllables presented dichotically with the first formant (F1) on one channel and the second and third formants (F2,F3) on the other channel was compared with perception of the same syllables presented binaurally. The syllables [ba, da, ga] were produced with the Haskins Laboratories parallel resonance synthesizer and digitized. A revised version of the Haskins Laboratories pulse code modulation (PCM) system (Cooper and Mattingly, 1969) was used to prepare audio tapes. Separate tapes were recorded for the dichotic experimental condition and the binaural condition. On the dichotic tape, F2,F3 of each syllable was recorded at six levels of attenuation, ranging from 0 to 50 db in 10-db steps. On half of the trials, F1 was recorded on channel 1 and F2,F3 on channel 2; on the remaining trials, this relationship was reversed. With five repetitions of each trial, there were 180 trials: (3 consonants) x (2 ear/formant relationships) x (6 F2,F3 intensity levels) x (5 repetitions). The trials were recorded in a randomized order with four seconds between trials.

The binaural control tape used six levels of F2,F3 attenuation ranging from 0 to 30 db in 6-db steps. The attenuated F2,F3 signal was mixed with F1 and the composite was recorded on both channels. With six repetitions of each trial

---

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

there were 108 trials: (3 consonants) x (6 F2,F3 intensity levels) x (6 repetitions). Since the magnitude of the masking phenomenon had been estimated during pilot work, the unequal attenuation step size between the dichotic and binaural tapes was used in an attempt to cover the intensity ranges most effectively.

The experimental dichotic condition is illustrated in Figure 1a; Figure 1b shows the corresponding binaural control condition. The attenuator in the path leading from F2,F3 permits control over the relative intensities of these formants. Since the three syllables used in this experiment (Figure 2) differ acoustically only in the F2,F3 region, attenuating these formants tends to obscure the distinctiveness of the syllables.

Calibration signals were also recorded on both channels of both tapes. The calibration signal was a sustained [a], the vowel used in the syllables. This calibration signal served a number of purposes: first, it enabled the channels to be equalized for level when the tapes were played to listeners; second, it permitted presentation levels to be equated between the dichotic and binaural tapes; and third, it was used to measure the absolute presentation level at the earphones. In this way, the playback level was adjusted so that the presentation level of the vowel [a], and hence of the vocalic portion of the syllables, was 70 db SPL.

Four subjects (young college adults with normal hearing) heard both tapes at one session. Their task was to write down "b," "d," or "g" for each trial.

The results are shown in Figure 3, where percent correct responses are plotted against the various F2,F3 attenuation levels for both the binaural control condition (broken line) and the dichotic experimental condition (solid line). In both cases, performance is near 100% for small attenuations and decreases at higher attenuation levels. The high degree of overall identification performance indicates that fusion of the signals in the two channels took place. For the range of attenuations used, performance remains always above the chance level. Inspection of Figure 3 reveals a rather dramatic release from masking. Approximately 20 db of attenuation separates the two curves for equal performance levels of 90% or less, the dichotic condition permitting greater attenuation. This may be interpreted as evidence that the normal binaural mode involves a certain degree of masking of the higher formants and that presenting the stimuli dichotically results in a release from this masking. Thus the masking level difference (MLD) is on the order of 20 db.

EXPERIMENT II

The second experiment was similar to the first, the only difference being the way in which the syllables were formed from complementary pieces. Rather than simply separating the formants and leading them to separate channels, the F2,F3 transition that is the minimal acoustic segment differentiating the syllables was separated from the remainder. This remainder, all of F1 plus the steady-state portions of F2 and F3, was constant across all three syllables. These acoustic segments are displayed in Figure 4.

Tapes were prepared and the experiment was run as described for Experiment I, with the same group of listeners. The results are plotted in Figure 5. From a comparison of Figure 5 with Figure 3, it is apparent that a similar release from

Figure I

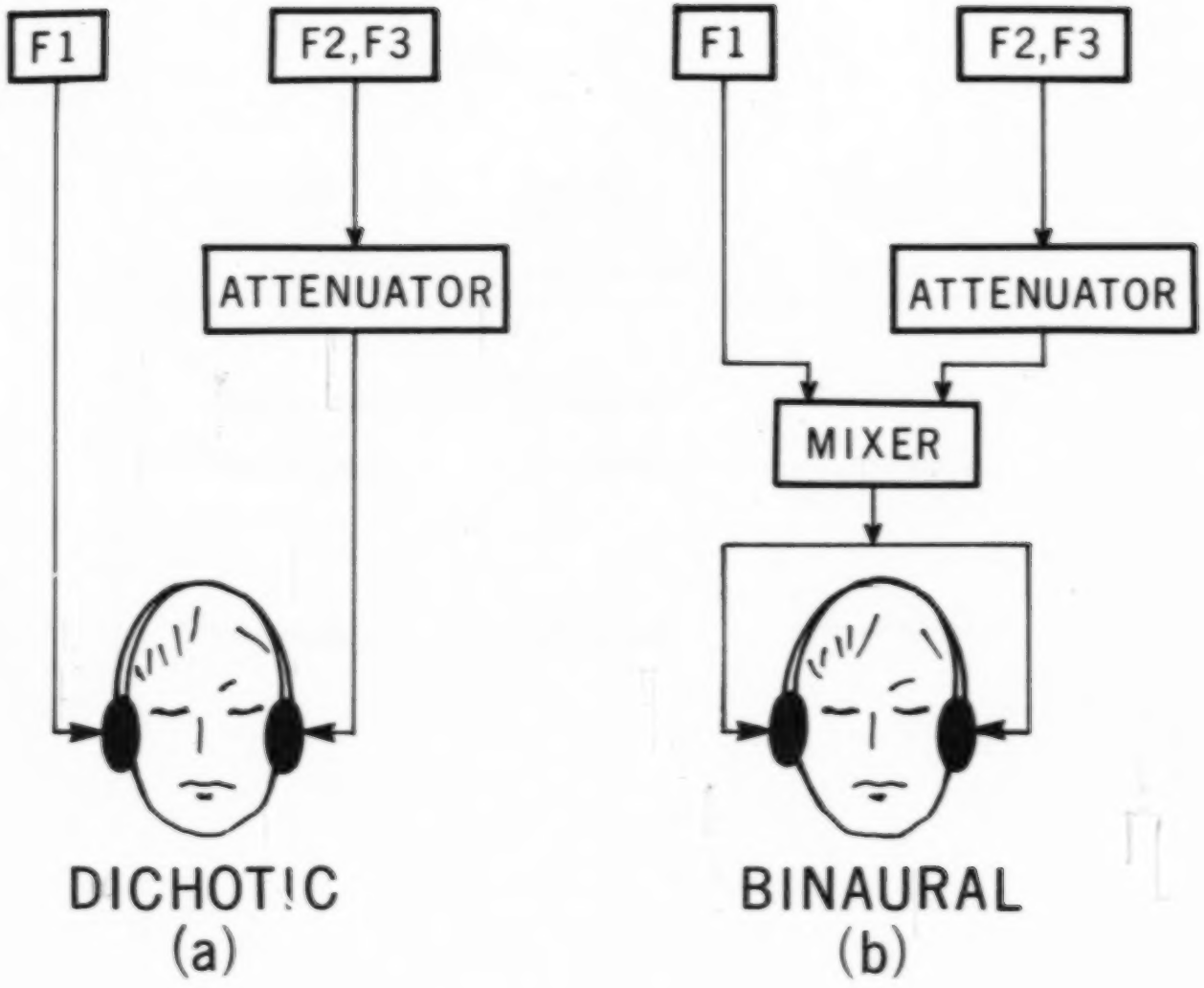
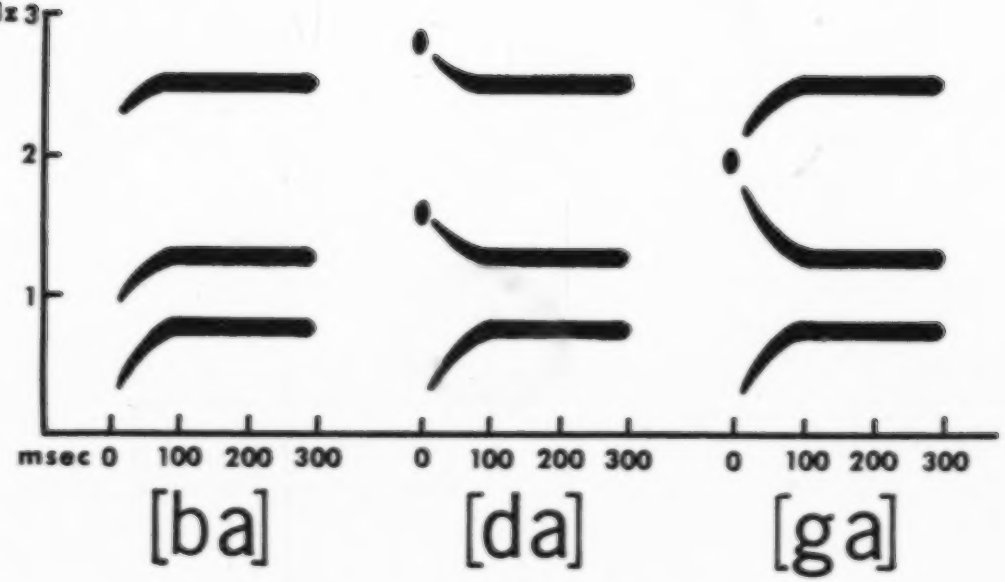

F1
F2,F3
ATTENUATOR
DICHOTIC
(a)
F1
F2,F3
ATTENUATOR
MIXER
BINAURAL
(b)

Figure 2



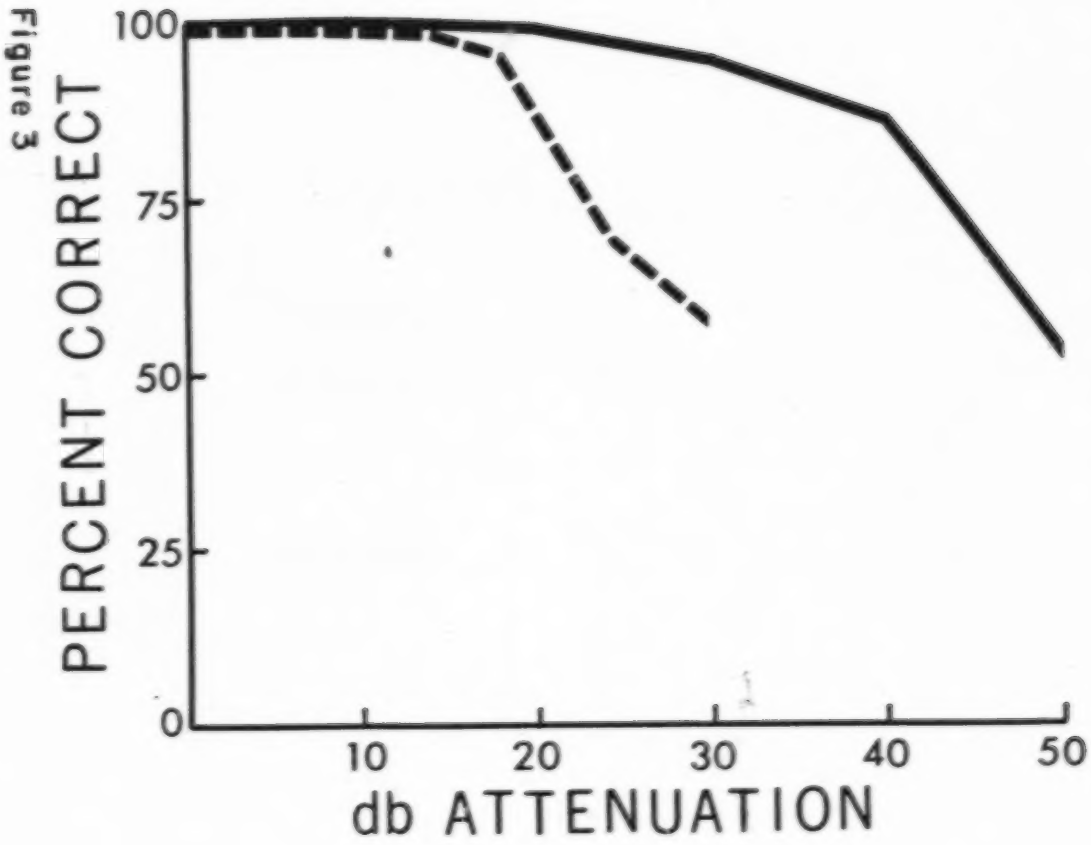


Figure 3: Experiment I results. Dichotic ———, binaural – – – –.

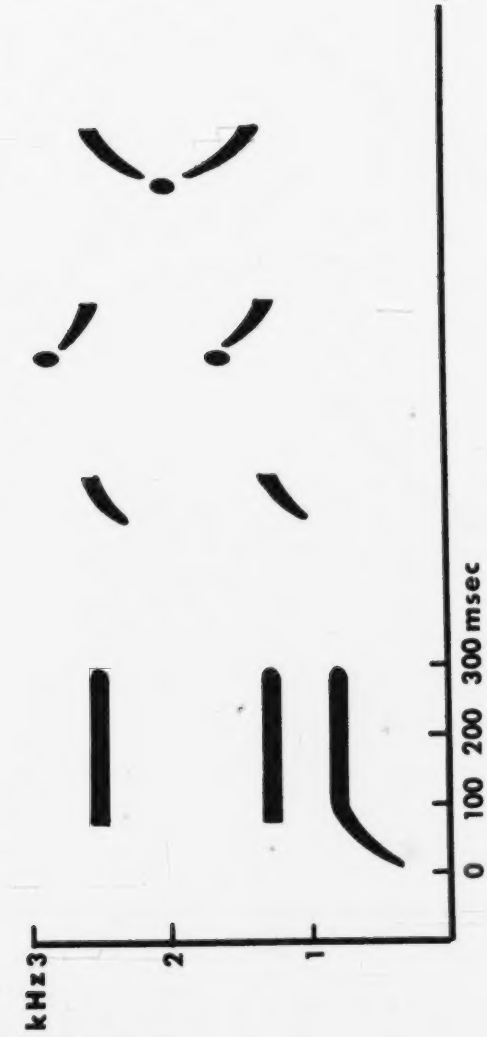


Figure 4

Figure 5

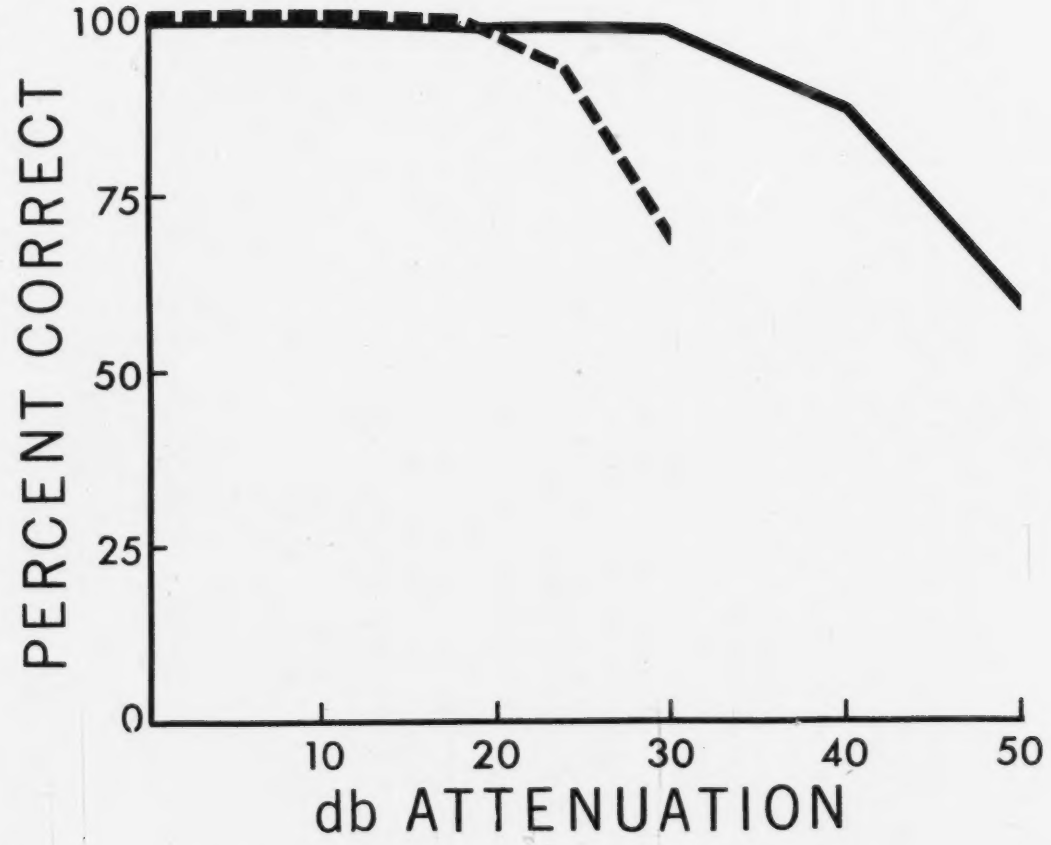


Figure 5: Experiment II results. Dichotic ———, binaural -----.

masking took place. The amount of this release, the MLD, is somewhat less for Experiment II (approximately 15 db).

Closer inspection of the two figures reveals that the primary dissimilarity is between the binaural conditions (broken lines). This says, in effect, that binaural presentation of the Experiment I stimuli, where F2,F3 were attenuated over the entire syllable, produces roughly 5 db greater masking than binaural presentation of Experiment II stimuli, where only the F2,F3 transitions were attenuated. For both experiments, dichotic presentation reduces this masking to a uniform minimum.

DISCUSSION

There is a certain practical significance to these results, particularly in the case of Experiment I. The release from masking achieved with synthetic speech suggests that a parallel gain could be achieved with natural speech were high and low band filtering combined with dichotic presentation. Any intelligibility gains by this method could prove to be of benefit in poor signal-to-noise environments or with hearing impaired listeners.

Experiment II, which divides the syllables both spectrally and temporally, serves as one step towards delineating the conditions under which signals reaching the two ears will or will not cooperate towards producing a unified percept. Spectral fusion is readily obtainable for speech, as has been noted by several investigators. Temporal fusion seems to have a different status. This is indicated, for example, by the work of Huggins (1964) on switching rates with continuous speech. In general, performance is poorest when speech switches between the ears at a rate of approximately once per syllable.

The stimuli of Experiment II produce an interesting phenomenon that bears on the question of speech/nonspeech processing. When presented with an F2,F3 transition in one ear and the remainder of a syllable in the other, listeners report hearing the syllable as well as a nonspeech sound. The nonspeech sound is heard at the ear to which, in fact, the F2,F3 transition is presented, and the syllable is heard as entering the other ear. Transitions of this type are generally heard as speech events or auditory events depending upon the context in which they occur (Mattingly, Liberman, Syrdal, and Halwes, 1971). The Experiment II stimuli, representing a situation in which a speech "context" is presented to the contralateral ear, produce the experience of hearing both kinds of event at the same time. Liberman, Mattingly, and Turvey (1972) advanced this phenomenon as support for the notion that the perceptual distinction between speech and nonspeech is not made at some early stage on the basis of broad acoustic characteristics.

REFERENCES

Broadbent, D. E. (1955) A note on binaural fusion. Quart. J. Exp. Psychol. 7, 46-47.

Broadbent, D. E. and P. Ladefoged. (1957) On the fusion of sounds reaching different sense organs. J. Acoust. Soc. Amer. 29, 708-710.

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. Haskins Laboratories Status Report on Speech Research SR-17/18, 17-21.

Flanagan, J. L. and M. G. Saslow. (1958) Pitch discrimination for synthetic vowels. J. Acoust. Soc. Amer. 30, 435-442.

Halwes, T. G. (1969) Effects of dichotic fusion on the perception of speech. Unpublished Ph.D. thesis, University of Minnesota. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Huggins, A. W. F. (1964) Distortion of the temporal pattern of speech: interruption and alternation. J. Acoust. Soc. Amer. 36, 1055-1064.

Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington, D. C.: V. H. Winston) 307-334.

Mattingly, I. G., A. M. Liberman, A. K. Syrdal, and T. Halwes. (1971) Discrimination in speech and nonspeech modes. Cog. Psychol. 2, 131-157.

Speech Misperception: Inferences About a Cue for Cluster Perception from a Phonological Fusion Task

James E. Cutting[+]
Haskins Laboratories, New Haven, Conn.

The acoustic cues for most speech sounds have been extensively researched. In fact, we know the contextual rules of acoustic phonetics well enough so that computer synthesis of speech from phonetic inputs is not only eminently feasible but a pleasant reality (see Liberman, Ingeman, Lisker, Delattre, and Cooper, 1959; Holmes, Mattingly, and Shearme, 1964). Yet some speech sounds remain difficult to synthesize by rule given our present knowledge. One of the most difficult speech synthesis problems is the production of high quality consonant clusters. Initial stop consonant plus liquid clusters are particularly difficult to synthesize well in all vowel contexts.

An interesting anomaly which occurs in a phonological fusion task may be pertinent to the discussion of cues in cluster perception. Cutting and Day (1972) found that dichotic pairs consisting of compatible stop-initial and liquid-initial stimuli yielded unusual results. While the dichotic pair PAY/LAY typically yielded a PLAY response, the dichotic pair PAY/RAY also yielded a PLAY response. Indeed, this misperception, the substitution of /l/ for /r/, occurred much more frequently than the more appropriate response PRAY; nearly all fusions were PLAY regardless of which liquid-initial stimulus was paired with PAY. This pattern of results has been found in phonological fusion tasks using natural speech stimuli (Day, 1968; Cutting, 1973) and synthetic speech stimuli (Cutting and Day, 1972; Cutting, 1973, in preparation-a). The results of these studies indicate that the acoustic cues which are crucial for the discrimination of /r/ versus /l/ are not relevant to the perception of stop + /r/ and stop + /l/ in the fusion task.

Perhaps misperceptions in the fusion task can provide information about important acoustic cues in the perception of stop/liquid clusters. We know that misperceptions occur very frequently in the fusion task, but we do not know their cause. More information is needed about what the subject perceives, given the stimuli PAY/LAY and PAY/RAY. Do both pairs sound like PLAY; or do the pairs sound different, and PLAY is just the best label for both? To aid discussion of these alternative hypotheses the results of four separate tasks are presented. The first two tasks were designed to determine the identifiability and the discriminability of the liquid-initial stimuli presented by themselves. A third task, a fusion task, was included to insure that the subjects tested misperceived

---

[+]Also Yale University, New Haven, Conn.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

stop + /r/-initial pairs. Finally, a discrimination task was devised to test the discriminability of dichotic pairs such as PAY/LAY and PAY/RAY.

Stimuli. Four sets of stimuli of the same general pattern were selected: the PAY set (PAY, RAY, LAY), the BED set (BED, RED, LED), the CAM set (CAM, RAM, LAMB), and the GO set (GO, ROW, LOW). Each stimulus was synthesized on the Haskins Laboratories parallel resonance synthesizer and transferred to the pulse code modulation (PCM) system (Cooper and Mattingly, 1969) for the preparation of diotic and dichotic tapes. These stimuli had been used previously by Cutting and Day (1972) and Cutting (in preparation-a) in other dichotic fusion tasks. The stimuli in each set were identical in pitch, intensity, and duration; and differed only in the acoustic structure of the initial 150 msec. The liquid-initial stimuli in each set (e.g., RAY and LAY) differed only in the direction and extent of the third formant transition, the acoustic cue most relevant for discrimination of liquids (O'Connor, Gerstman, Liberman, Delattre, and Cooper, 1957; Lisker, 1957).

Subjects. Twelve Yale University undergraduates received course credit for their participation in four tasks. Each subject was right-handed and a native American English speaker who had no history of hearing difficulty.

Procedure. Subjects were tested in groups of four. They listened to tapes played on an Ampex AG500 dual track tape recorder, sent through a listening station of Grason Stadler earphones (model TDH39-300Z). All subjects participated in four tasks in the order: a) dichotic fusion; b) dichotic fusible-pair discrimination; c) diotic liquid identification; and d) diotic liquid discrimination.

In dichotic tasks two different stimuli are presented, one to each ear; in diotic tasks the same stimulus is presented to both ears at the same time. Subjects hear a diffuse and ambiguous mixture of stimuli in the dichotic task and they may perceive one item or two items. In diotic tasks, on the other hand, they hear one clear item localized at the midline. All subjects participated in the dichotic tasks first so that specific information about the stimuli that they would gain in the diotic tasks could not influence their perceptions in the dichotic situation. In considering these tasks, however, it is more fruitful to discuss the diotic tasks before the dichotic tasks. Each task is discussed more fully in turn.

TASK 1 - LIQUID IDENTIFICATION

A brief test was run to assess the identifiability of the liquid stimuli.

Tape and procedure. A diotic identification tape was prepared consisting of 48 liquid-initial items presented one at a time: (4 sets of stimuli) x (2 liquids per set) x (6 observations per stimulus). There was a three-second interval between liquid-initial items. Subjects wrote down the first sound that they heard in each stimulus, choosing between the liquids /r/ and /l/.

Results. All liquid-initial stimuli were highly identifiable. The /r/-initial stimuli were correctly identified on 88 percent of all trials, while the /l/-initial stimuli were identified correctly on 91 percent of all trials. Previous testing determined that the stop-initial stimuli were also very identifiable.

TASK 2 - LIQUID DISCRIMINATION

A second test was run on the liquid-initial stimuli to assess their discriminability.

Tape and procedure. A brief diotic liquid discrimination tape was prepared. It consisted of liquid-initial stimulus triads, such as LAY-RAY-LAY as shown in the top of Figure 1, with 1 second between members of the triad and 4 seconds between triads. It was a typical ABX procedure.[1] The subjects were told that they would hear three stimuli per trial, and that they should listen carefully for the differences between the first stimulus (Stimulus A) and the second stimulus (Stimulus B). The third stimulus (Stimulus X) was an identical match of either Stimulus A or Stimulus B. The subjects' task was to decide which stimulus was the match. All four possible ABX arrangements for each set of liquids were used: RLR,[2] RLL, LRL, LRR. There were 96 ABX triads: (4 sets of stimuli) x (4 ABX arrangements per set) x (6 observations per ABX triad).

Results. Subjects had little difficulty discriminating the liquids. Figure 2 shows that the mean discrimination rate for all stimuli and for all subjects was 80 percent, with a range from 73 to 95 percent. Thus, the liquid-initial stimuli were not only easily identifiable but also easily discriminable.

TASK 3 - PHONOLOGICAL FUSION

A fusion task was run to insure that these subjects misperceived stop + /r/ stimulus pairs as previous subjects have done.

Stimuli and procedure. Dichotic pairs were assembled from stop-initial and liquid-initial stimuli within the same set, for example PAY/RAY and PAY/LAY. Note that all stimuli and all possible fusions are high frequency monosyllabic words; among others, PAY/RAY⟶PRAY and PAY/LAY⟶PLAY.[3] The members of each pair were recorded at three relative onset times: they began at the same time, or one stimulus began 50 msec before the other. LAY began before PAY and PAY before LAY on an equal number of trials. Channel arrangements for items within a dichotic pair were counterbalanced: for example, on half the trials PAY was recorded on channel 1 of the audio tape and LAY was recorded on channel 2, while the reverse configuration was recorded on the other half of the trials. A dichotic fusion tape of 96 items was prepared; (4 sets of stimuli) x (2 stop/liquid pairs per set) x (3 lead times) x (2 channel arrangements) x (2 observations per pair). Subjects wrote down what they heard; they were given the alternatives of writing down real words or nonsense words, one word or two, for each presentation.

Results and preliminary discussion. Fusion occurred readily for all stimulus pairs: subjects fused on 64 percent of all trials. Fusion rates were roughly

---

[1] For a discussion of the relative merits of the ABX procedure as opposed to other discrimination procedures, see Pisoni (1971).

[2] For /r/-/l/-/r/ as the A-B-X stimuli respectively.

[3] The arrow(⟶) should be read as "yields."

Figure 1

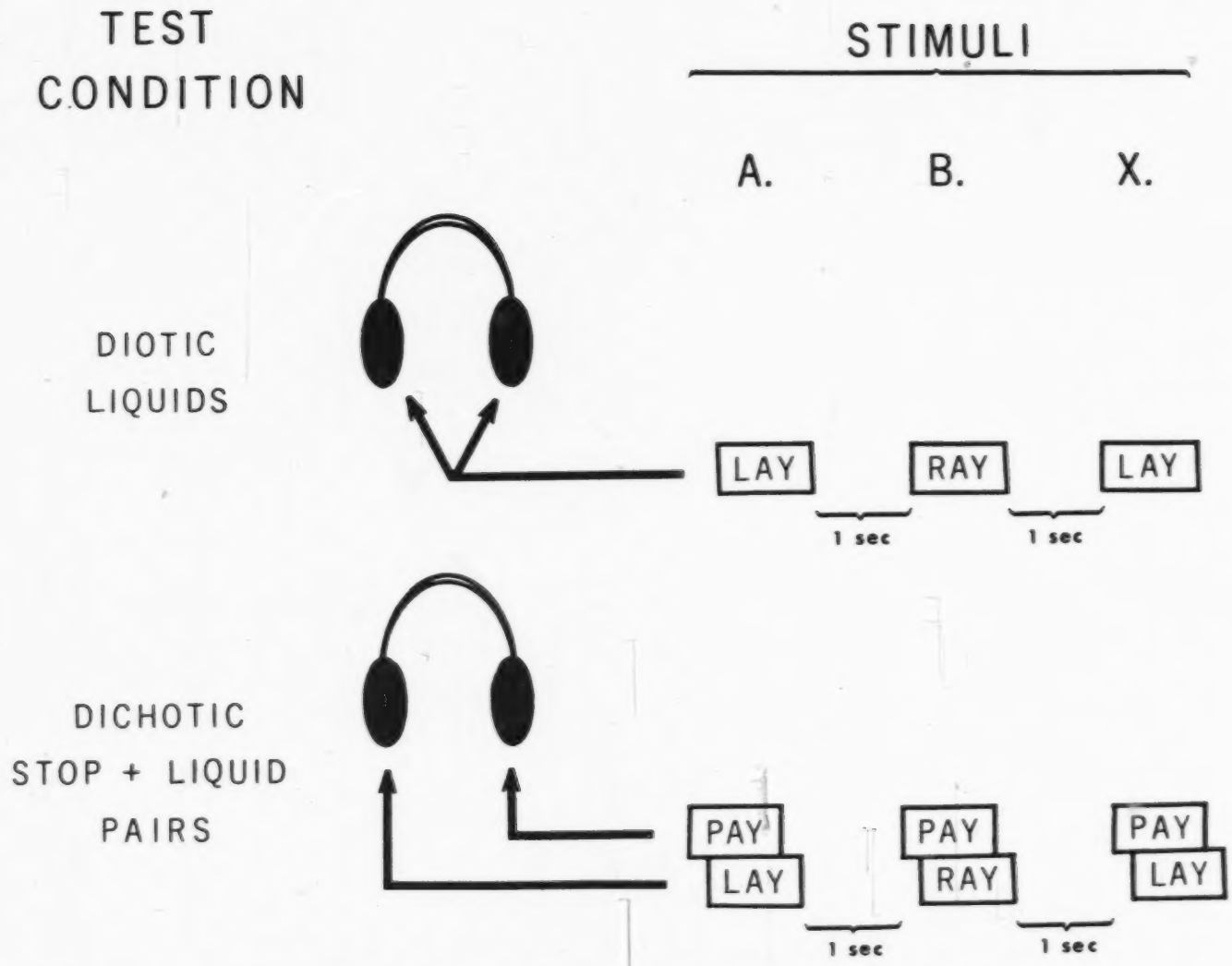


Figure 1: The general paradigm of diotic (Task 2) and dichotic (Task 4) discrimination tasks.

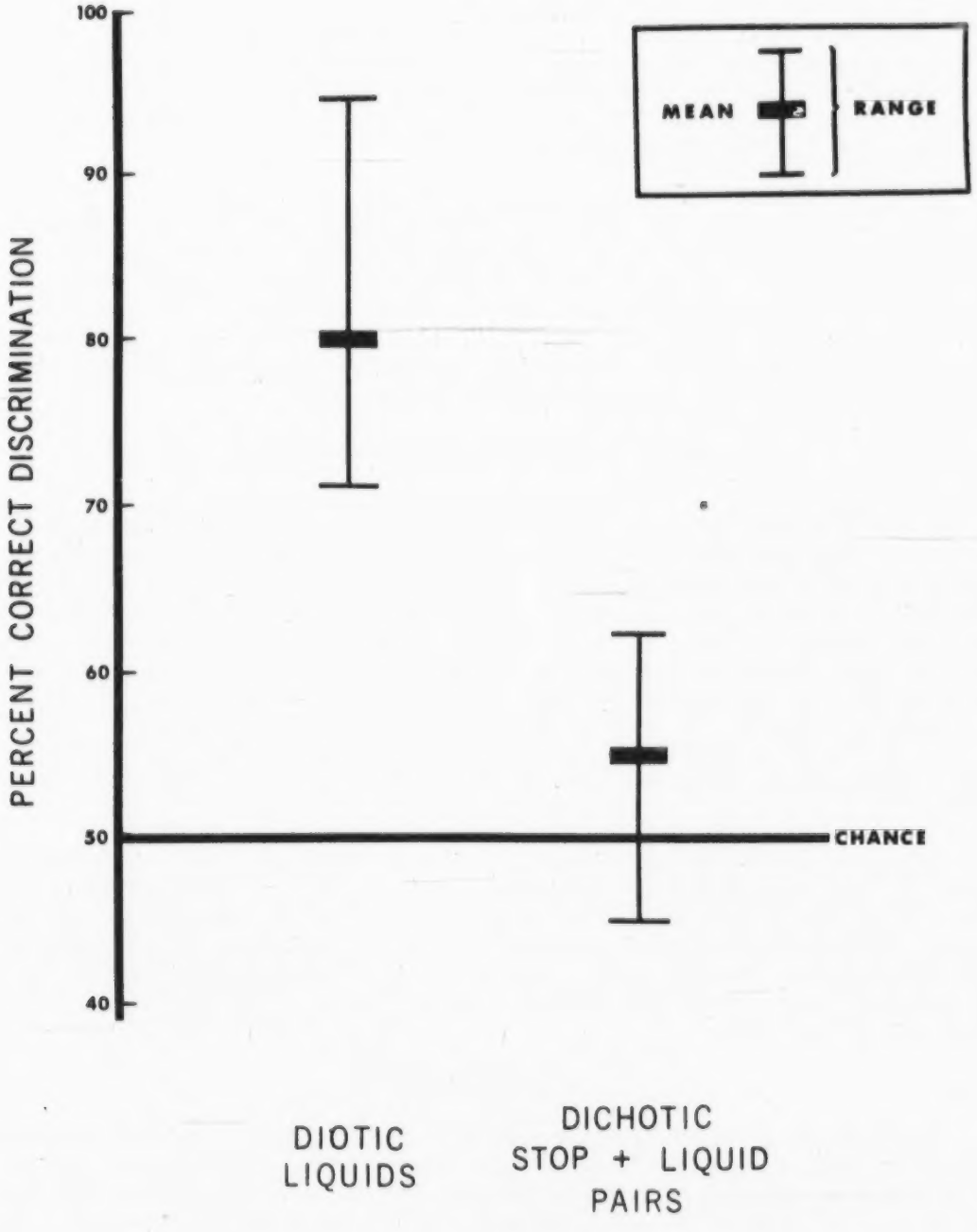


Figure 2: Mean and ranges for the two discrimination tasks.

comparable for all sets of stimuli Lead times and channel arrangements were not significant factors.

Fusion responses showed the pattern expected from the results of previous studies. Almost all fusion responses were stop + /l/ items regardless of the stimuli presented. Consider once again the PAY set. Both PAY/LAY and PAY/RAY pairs yielded the fusion PLAY. The other sets showed this same pattern: an /r/-initial stimulus paired with an appropriate stop-initial stimulus yielded a stop + /l/ fusion. This /l/-for-/r/ substitution occurred on 78 percent of all trials in which stop + /r/ stimuli fused. The reverse substitution rarely occurred. These results cannot be accounted for by the relative frequency of occurrence of these clusters in English. In fact, stop + /r/ clusters outnumber stop + /l/ clusters by a significant margin (Day, 1968).

Since we have found similar results in previous studies, the /l/-for-/r/ substitutions in the fusion task were anticipated. Therefore, a fourth task was designed to probe the reason for this misperception. Two alternative explanations are possible: a) stop + /r/ and stop + /l/ stimuli sound identical and therefore subjects give them the same label; or b) stop + /r/ pairs sound different from stop + /l/ pairs but since the difference cannot be labeled subjects assign both the same name. The fourth task poses the question: are stop + /r/ and stop + /l/ pairs of the same set discriminable?

TASK 4 - FUSIBLE-PAIR DISCRIMINATION

Tape and procedure. A modified ABX procedure was designed. As shown in the bottom part of Figure 1, each ABX triad was composed of three dichotic pairs, such as PAY/LAY-PAY/RAY-PAY/LAY. There was one second between fusible pairs within the triad and four seconds between triads. The stop-initial stimulus in each of the three pairs was identical, as was the channel assignment and the lead time configuration. Only the liquid stimuli were varied within the triad, and they were varied in the same manner as in Task 2. Subjects wrote A or B as their choice for the member of the triad which matched stimulus X. The tape was 96 trials in length: (4 sets of stimuli) x (4 ABX arrangements per stop/liquid set) x (3 lead time configurations) x (2 channel assignments per ABX triad). Lead time configurations were the same as those used in Task 3, and were counterbalanced.

Results. Subjects had extreme difficulty with the task. As shown in Figure 2, they averaged only 54 percent correct, with a range of 46 to 63 percent. Only two of the 12 subjects scored significantly above chance, and all subjects scored considerably worse on this task (by at least 20 percentage points) than on the liquid discrimination task. Moreover, the distributions for the two tasks did not overlap. Lead times and channel assignments were not significant factors affecting the results.

DISCUSSION

The results of Tasks 2 and 4 show that while subjects can discriminate liquid-initial stimuli presented singly, they cannot discriminate them in dichotic fusible pairs. The third formant transition, the most potent cue for the discrimination of /r/ versus /l/, is irrelevant in the dichotic fusion case. Some other cue, as yet undefined, overrides the effect of the third-formant transition and

the resulting fusions are items containing stop + /l/ clusters. The /l/-for-/r/ substitutions in the fusion task appear to be based on identical perceptions by the subjects, not on similar perceptions given the same label: PAY/LAY⟶PLAY and PAY/RAY⟶PLAY, and subjects cannot tell the difference between them. The question arises: what is the cue for stop + /l/ clusters in the fusion task, and where does the cue come from?

Figure 3 shows natural speech productions of PLAY and PRAY recorded by the author. Consider the differences between the two utterances, especially between the end of aspiration in the /p/ segment and the onset of the diphthong /ɛI/. PLAY has a segment of low amplitude steady-state formant structure after the /p/ and before the release of the /l/. The duration of this segment is about 40 msec. PRAY has no corresponding segment; the formant transitions of the /r/ appear to rise directly after the aspiration of the /p/. In PLAY the boundary between the 40 msec steady-state segment and the rise of the formants in the /l/ marks a discontinuity in both formant structure and intensity. This discontinuity is not present in PRAY. Perhaps this break in the acoustic pattern is an important cue for the perception of stop + /l/ clusters, a cue which may serve to distinguish stop + /l/ clusters from stop + /r/ clusters.

Now reconsider the dichotic fusion task. When subjects listen to a stop-initial stimulus presented to one ear and a compatible liquid-initial stimulus presented to the other, they listen to an unusual combination of sounds. Indeed, on practice trials which preceded the fusion task subjects were often quite perplexed by what they heard. Typically they needed reassurance that most subjects thought these items sounded rather peculiar, but that they should perform the task as best they could. Perhaps the dichotic presentation itself provides a cue for the perception of stop + /l/ clusters. This presentation may set up a kind of discontinuity of a spatial nature which may serve as a surrogate cue replacing the acoustic and intensity discontinuities in the naturally occurring stop + /l/ cluster.

What happens when the dichotic cues of discontinuity are removed? Another experiment used the same fusible stimuli as in the present series of tasks, but presented them in two modes (Cutting, in preparation-b). One mode was dichotic and the other was binaural. In the dichotic mode different stimuli were presented to opposite ears: for example, PAY was presented to one ear and LAY to the other. In the binaural mode, however, the stimuli were electrically mixed and both were presented to both ears: thus, PAY and LAY were presented in their acoustic entirety to both ears.

The binaural presentation mode was identical to the dichotic mode in all respects except that the cue of spatial discontinuity was removed. Fusion levels were observed for both modes. The fusion rate was markedly higher in the dichotic condition than in the binaural condition, with a ratio of greater than 3/1. As usual, nearly all fusions in the dichotic condition were stop + /l/ clusters regardless of the stimuli presented. Differences in fusion rates can only be accounted for by the presentation modes. Fusions occurred quite readily in the dichotic mode when the spatial discontinuity was present in the fusible pair. Fusions did not occur frequently when this cue was removed.

Given the appropriate stimuli, spatial discontinuity in the dichotic task may account for /l/-for-/r/ substitution. This cue appears to override the

NATURAL SPEECH

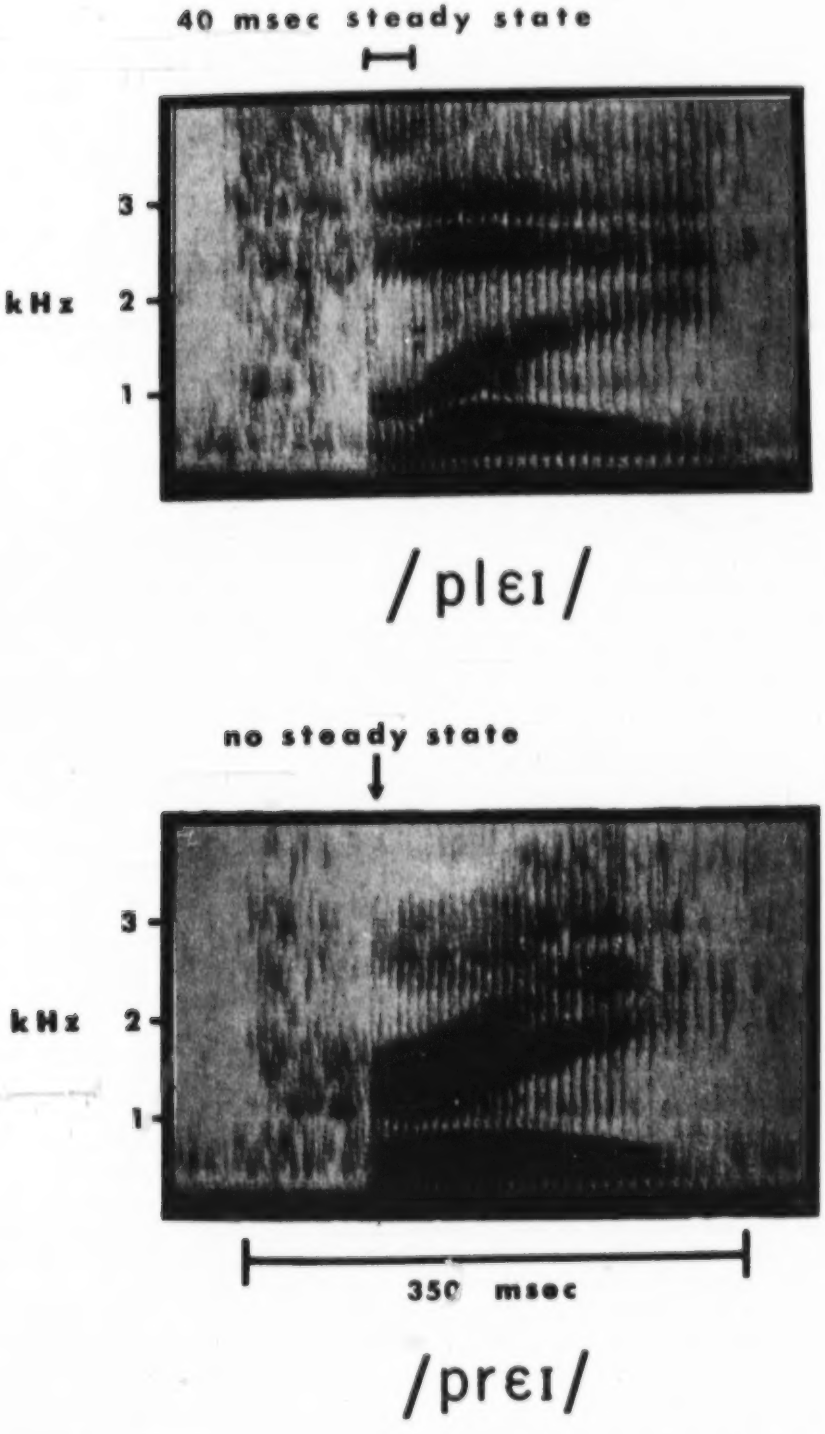


Figure 3: Spectrograms of natural speech productions of PLAY and PRAY.

acoustic cue of the third-formant transition in the liquid, a requisite for the perception of /r/ versus /l/, but not a requisite for the perception of stop + /r/ versus stop + /l/ in the phonological fusion task. Perhaps we have also found an important cue for the perception of naturally occurring stop/liquid clusters in English. Synthesizing good stop + liquid clusters for all vowel contexts in English is not an easy task given our present knowledge. Perhaps an additional rule is needed: discontinuity in formant structure. Stop + /l/ clusters appear to require this acoustic cue [+ discontinuity], while stop + /r/ clusters must not have it [- discontinuity].[4] The idea is provocative.

REFERENCES

Cooper, F. S. and I. G. Mattingly. (1969) Computer-controlled PCM system for investigation of dichotic speech perception. J. Acoust. Soc. Amer. 46, 115(A).

Cutting, J. E. (1973) Phonological fusion in synthetic and natural speech. Haskins Laboratories Status Report on Speech Research SR-33 (this issue).

Cutting, J. E. (in preparation-a) Levels of processing in phonological fusion. Unpublished Ph.D. thesis, Yale University (Psychology).

Cutting, J. E. (in preparation-b) Phonological fusion of dichotic and binaural stimuli.

Cutting, J. E. and R. S. Day (1972) Dichotic fusion along an acoustic continuum. J. Acoust. Soc. Amer. 52, 175(A). (Also in Haskins Laboratories Status Report on Speech Research SR-28.)

Day, R. S. (1968) Fusion in dichotic listening. Unpublished Ph.D. thesis, Stanford University (Psychology).

Holmes, J. N., I. G. Mattingly, and J. N. Shearme. (1964) Speech synthesis by rule. Language and Speech 7, July-Sept., 127-143.

Jakobson, R., C. G. M. Fant, and M. Halle. (1951) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).

Liberman, A. M., F. Ingeman, L. Lisker, P. C. Delattre, and F. S. Cooper. (1959) Minimal cues for synthesizing speech. J. Acoust. Soc. Amer. 31, 1490-1499.

Lisker, L. (1957) Minimal cues for separating /w, r, l, y/ in intervocalic position. Word 13, 257-267.

O'Connor, J. D., L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1957) Acoustic cues for the perception of initial /w, j, r, l/ in English. Word 13, 25-43.

Pisoni, D. B. (1971) On the nature of categorical perception of speech sounds. Unpublished Ph.D. thesis, University of Michigan (Psycholinguistics). (Issued as a Supplement to Haskins Laboratories Status Report on Speech Research.)

[4] The cue I have suggested here, however, is not the same as the continuant/discontinuant feature suggested by Jakobson, Fant, and Halle (1951).

Cross-Language Study of the Perception of the F3 Cue for [r] versus [l] in Speech- and Nonspeech-Like Patterns*

Kuniko Miyawaki,[+] A. M. Liberman,[++] O. Fujimura,[+++] Winifred Strange,[++++] and J. J. Jenkins[++++]

Our primary purpose in the experiments reported here was to see how linguistic experience affects the way we perceive the sounds of speech. To do that, we compare the discrimination of a particular phonetic contrast by two groups of people. The one group speaks a language which makes use of the distinction under study; the other group does not. If we find a difference in ability to discriminate, we attribute it to familiarity with the distinction, in the one case, and unfamiliarity in the other.

Several experiments employing this paradigm had been carried out before we undertook the one under discussion. Perhaps the first was a study on vowel discrimination by Stevens, Liberman, Studdert-Kennedy, and Öhman (1969). Using controlled synthetic patterns, these investigators found essentially no effect of linguistic experience: listeners who were unfamiliar with the vowels discriminated among them quite as well as those who used the vowels frequently in their native language. Studies on the discrimination of the voicing distinction in stop consonants have, however, yielded a different result. In that case, Abramson and Lisker (1970) found that listeners who were familiar with the distinction showed an increase in discrimination of the phonetic quality at the phonemic boundary; listeners who were unfamiliar with it did not.

That linguistic experience affects the perception of vowels and stop consonants differently is, perhaps, to be related to other differences between these two classes of sounds. Vowels are represented rather directly in the sound stream; they are, like most nonspeech sounds, perceived in continuous fashion; and, according to the results of some other experiments, they can apparently be processed in either cerebral hemisphere. Stop consonants, on the other hand, are complexly encoded in the sound stream; they are perceived in nearly categorical fashion; and they apparently need to be processed in the left (or language) hemisphere.

---

*Presented at the XXth International Congress of Psychology, Tokyo, 1972.

[+]Department of Linguistics, University of Tokyo, Japan.

[++]Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

[+++]Faculty of Medicine, University of Tokyo, Japan.

[++++]Center for Research in Human Learning, University of Minnesota, Minneapolis.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

The experiment reported here deals with the effect of experience on the discrimination of [r] and [l]. There are at least four reasons for studying this distinction. First, these two segments constitute a class of speech sounds that appear to be intermediate between the vowels and stop consonants in regard to the differences just described. Second, the distinction is a natural one for cooperative research by Japanese and American colleagues, since it is familiar to native speakers of English but notoriously unfamiliar to the Japanese. Third, this distinction has an advantage over those studied earlier because it permits us not only to compare our two groups of listeners on discrimination of speech, but also to see how they discriminate the essential acoustic cue when it is presented in isolation and not heard as speech. And, fourth, the data we obtain will help to fill a gap in our knowledge of speech perception, since relatively little experimental work has been done on the discrimination of [r] and [l].

Figure 1 shows the stimuli, the experimental variable, and the two conditions of the experiment. The speech stimuli are shown at the left. There you see spectrograms of synthetic syllables. The one at the bottom is a reasonable approximation to [ra]; the one at the top approximates [la]. These two patterns are identical except for the starting point and transition of the third formant. For [r], at the bottom, the third formant starts low and rises; for [l], at the top, it starts high and falls. In the other stimuli in the series, the third formant started at levels that sampled the range between the extreme [r] and [l] values shown in the figure. There were 13 stimuli in all.

Corresponding samples of the nonspeech stimuli are shown at the right. These nonspeech stimuli are the formants of the speech patterns presented now in isolation. As these third formants constitute the only differences among the speech stimuli, they are alone responsible for the perceived differences among the speech stimuli. But when they are presented alone in the nonspeech condition, they do not sound at all like speech.

Figure 2 shows the results obtained with a group of 12 American listeners, and sets the stage for the comparison with the Japanese. The graph at the top shows what happened when the 13 synthetic speech stimuli were presented singly, in random order, for absolute identification as [ra] or [la]. The graph is plotted as the percent of [r] responses against the position of the third formant, indicated simply by the numbers 1 to 13. We see that the American listeners did sort the stimuli in a fairly consistent way.

The graph at the bottom shows how the same subjects discriminated the speech stimuli. To measure discrimination, we arranged the synthetic stimuli in sequences of three, such that two members of each sequence were identical and one was different. The listeners' task was to say which one was different. Each listener was instructed, for that purpose, to use any differences he could hear. Let me illustrate a result. The first datum point at the left shows what happened when the stimuli in the triplet were, as indicated on the abscissa, Numbers 1 and 4. The point we see on the graph shows that our subjects correctly discriminated these two stimuli 40% of the time. All stimulus combinations were, like those, three steps apart on the stimulus scale.

Looking at the entire graph, we see that discrimination rose to a fairly high peak at a point corresponding approximately to the phonemic boundary, as shown in the identification function at the top. Thus, we have here an approximation to the kind of quantal or categorical perception that has been found

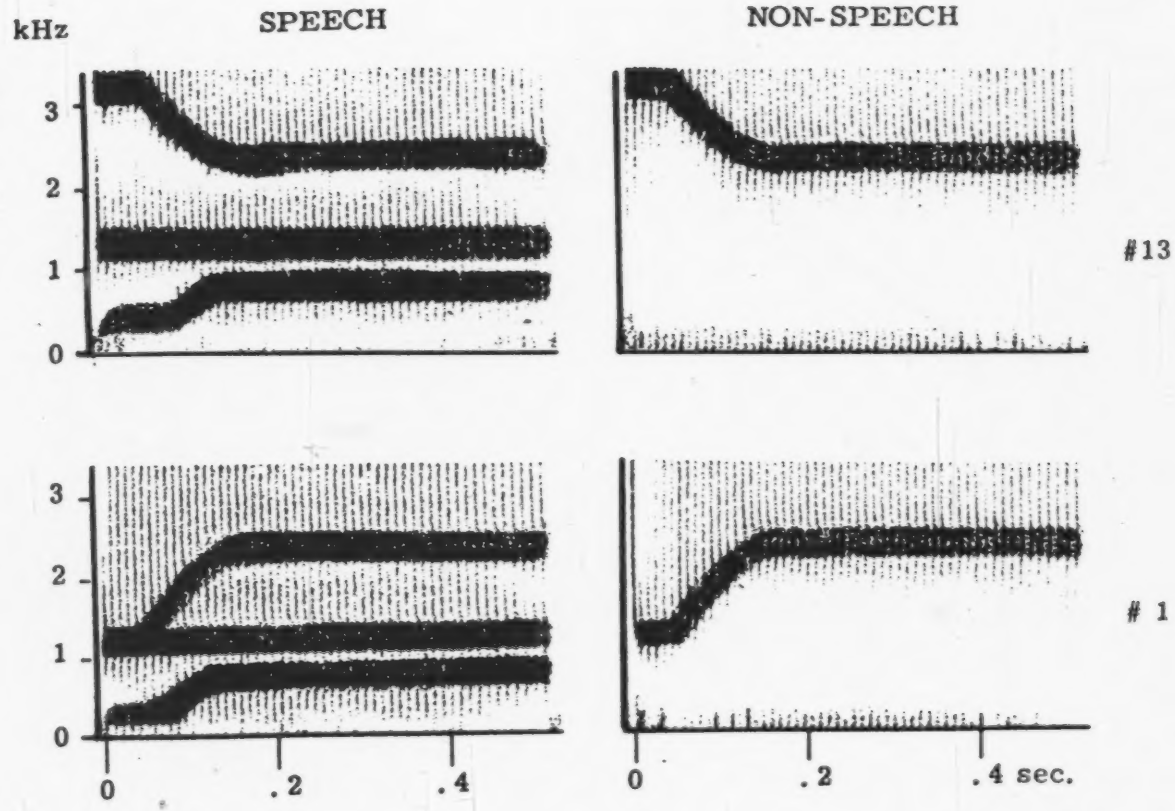


Figure 1: Spectrograms of speech and nonspeech stimuli--[la] upper and [ra] lower.

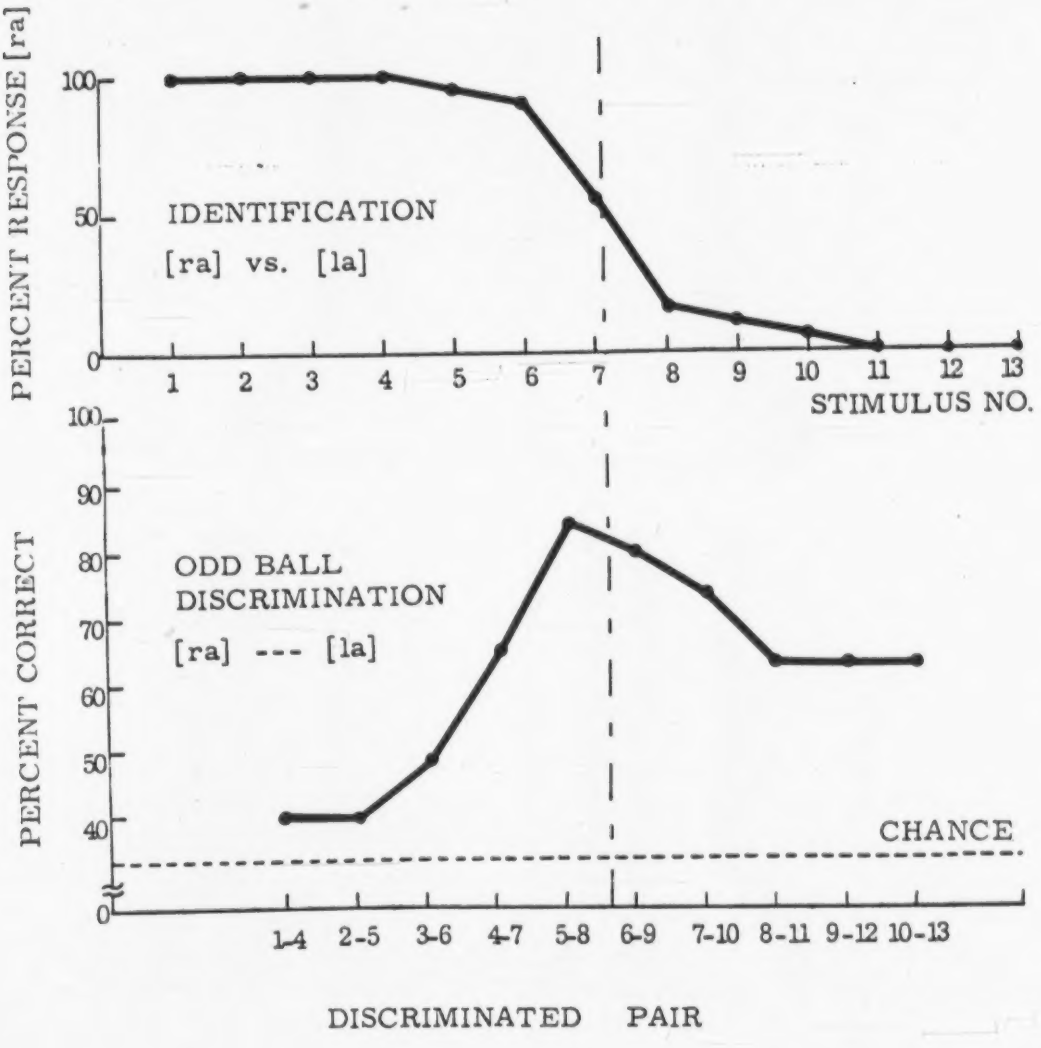


Figure 2: American results - identification vs. discrimination.

previously for the stops but not the vowels. We must emphasize, however, that a proper comparison with the categorical perception of the stops must await further research and analysis.

Figure 3 permits us to compare discrimination of the speech and nonspeech patterns. On the left we have reproduced the discrimination result, already shown in Figure 2, that was obtained with the speech-like stimuli. On the right are results obtained with the nonspeech stimuli--that is, with the third-formant pattern in isolation. We see that the nonspeech function is quite different from the speech function: the nonspeech function does not show the marked peak at the phoneme boundary and tends to be high throughout. We might infer, then, that the peaking at the phoneme boundary is not so much in the psychoacoustic response to the stimulus as it is a property of perception in the speech mode.

Figure 4 shows the results obtained with 21 Japanese listeners. Looking at the speech discrimination function at the left, we see that it is not peaked as the American function was, and is rather low throughout. The nonspeech function at the right shows that the Japanese subjects discriminated the third formant transition in isolation better than in the speech patterns with lower formants.

I should note that two of the Japanese subjects had spoken English as children. Their results were nevertheless included in the graph of Figure 4. In Figure 5 the data obtained with these two subjects are presented separately. We see that their discrimination functions look very much like those of the Americans'.

In Figure 6 we compare the overall results of the American and Japanese subjects. Looking first at the results with the speech-like syllables, at the left, we see that the Americans do indeed discriminate better than the Japanese, especially in the vicinity of the phoneme boundary. But the nonspeech functions at the right are remarkably alike. Thus, the two groups do not differ in their ability of discriminate the essential acoustic pattern when it is presented in isolation, but they do differ when the same pattern serves as a phonetic cue.

I will summarize. The acoustic cue responsible for the perceived distinction between [r] and [l] was, in one set of stimuli, varied in equal steps in synthetic syllables that could be perceived as [r] or [l] plus a vowel. In another set this acoustic characteristic was presented in isolation and did not sound like speech. American listeners, to whom the phonemic distinction is familiar, showed a peak in discrimination similar to those found in the discrimination of the stops. Discrimination of the crucial acoustic pattern in isolation was clearly different. Japanese listeners, who are generally quite unfamiliar with the [r-l] distinction, discriminated it very poorly by comparison with the Americans. Their relative inability to discriminate [r] and [l] cannot be attributed to an inability to discriminate the acoustic patterns, since their discrimination performance was at least equal to that of the Americans when the cue pattern was presented in isolation. Thus, we have here a rather large effect of linguistic experience on perception in the speech mode. In this respect, [r] and [l] are more like the stops than like the vowels. But whether the effect we found here is as large as that obtained with the stops can only be determined by further research.

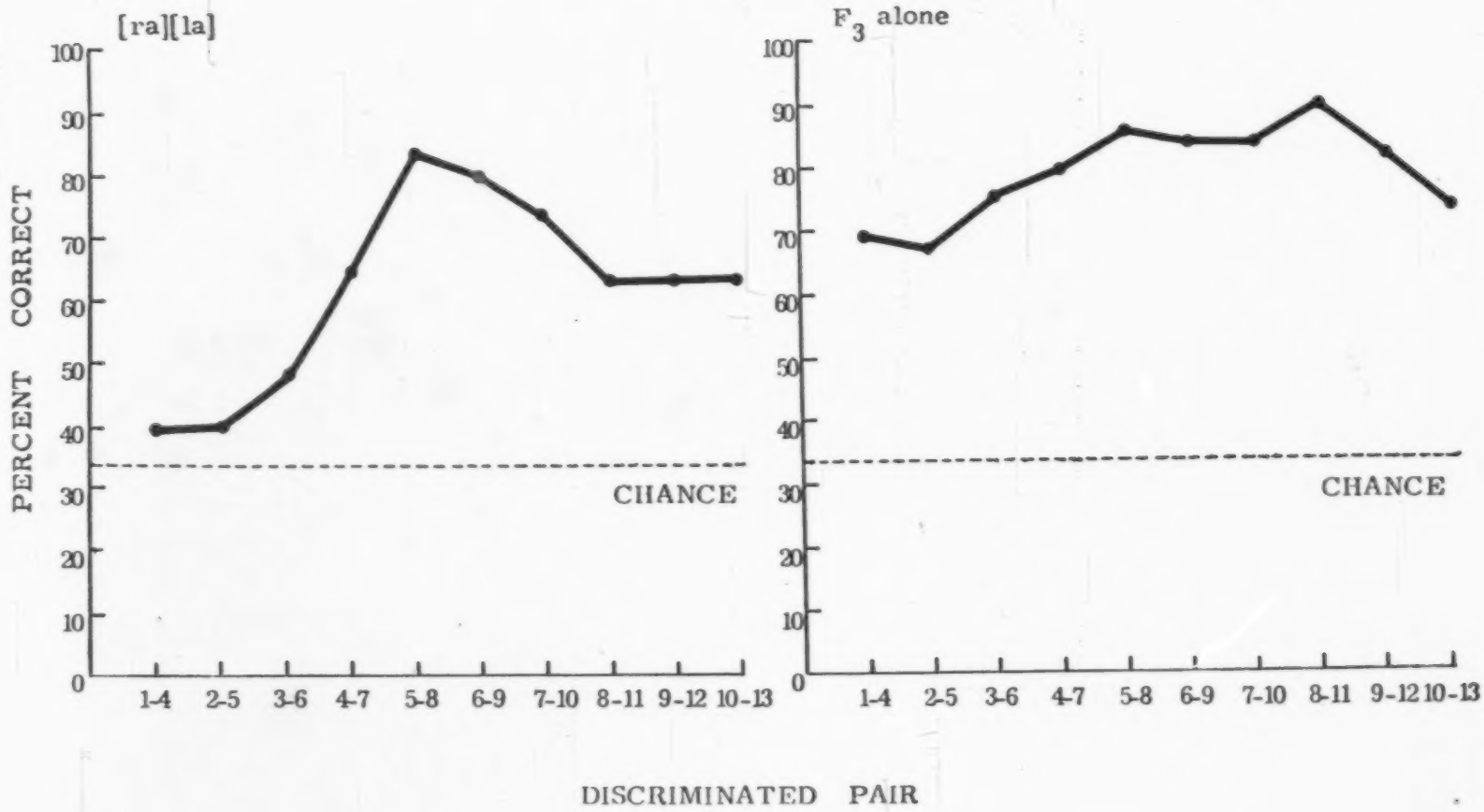


Figure 3

Figure 3: Overall American (13 subjects).

Figure 4

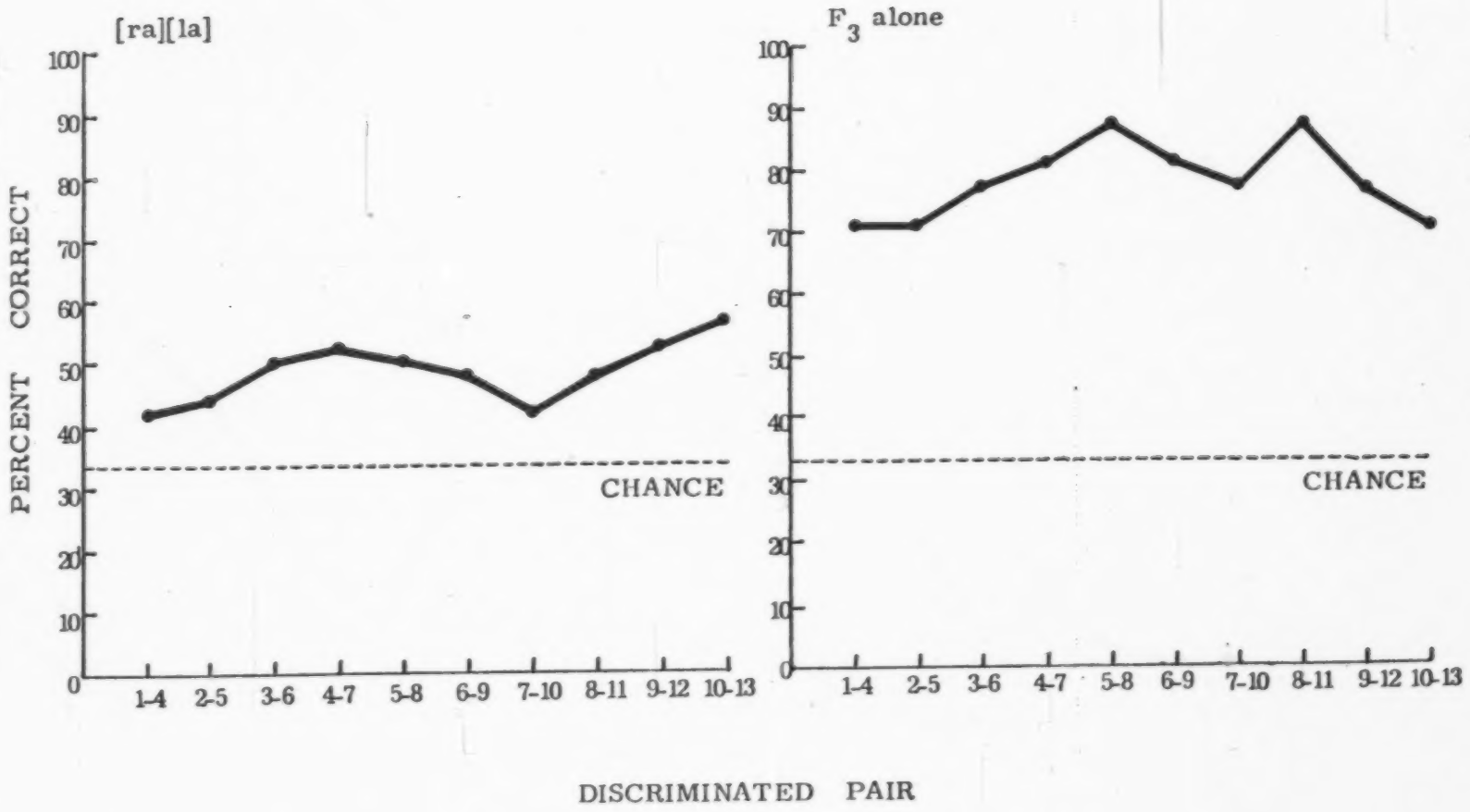


Figure 4: Overall Japanese (21 subjects).

Figure 5

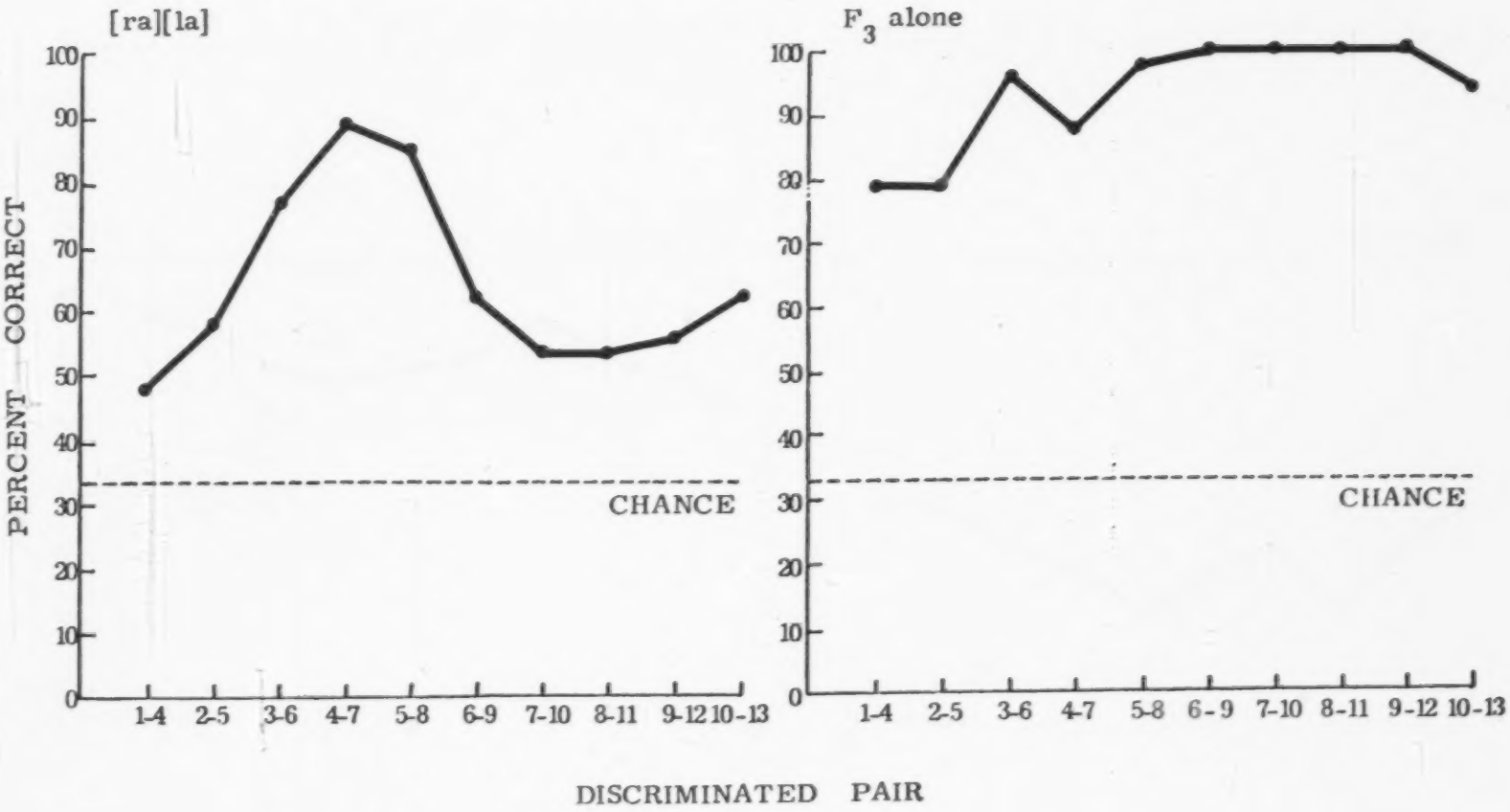


Figure 5: Two special Japanese subjects.

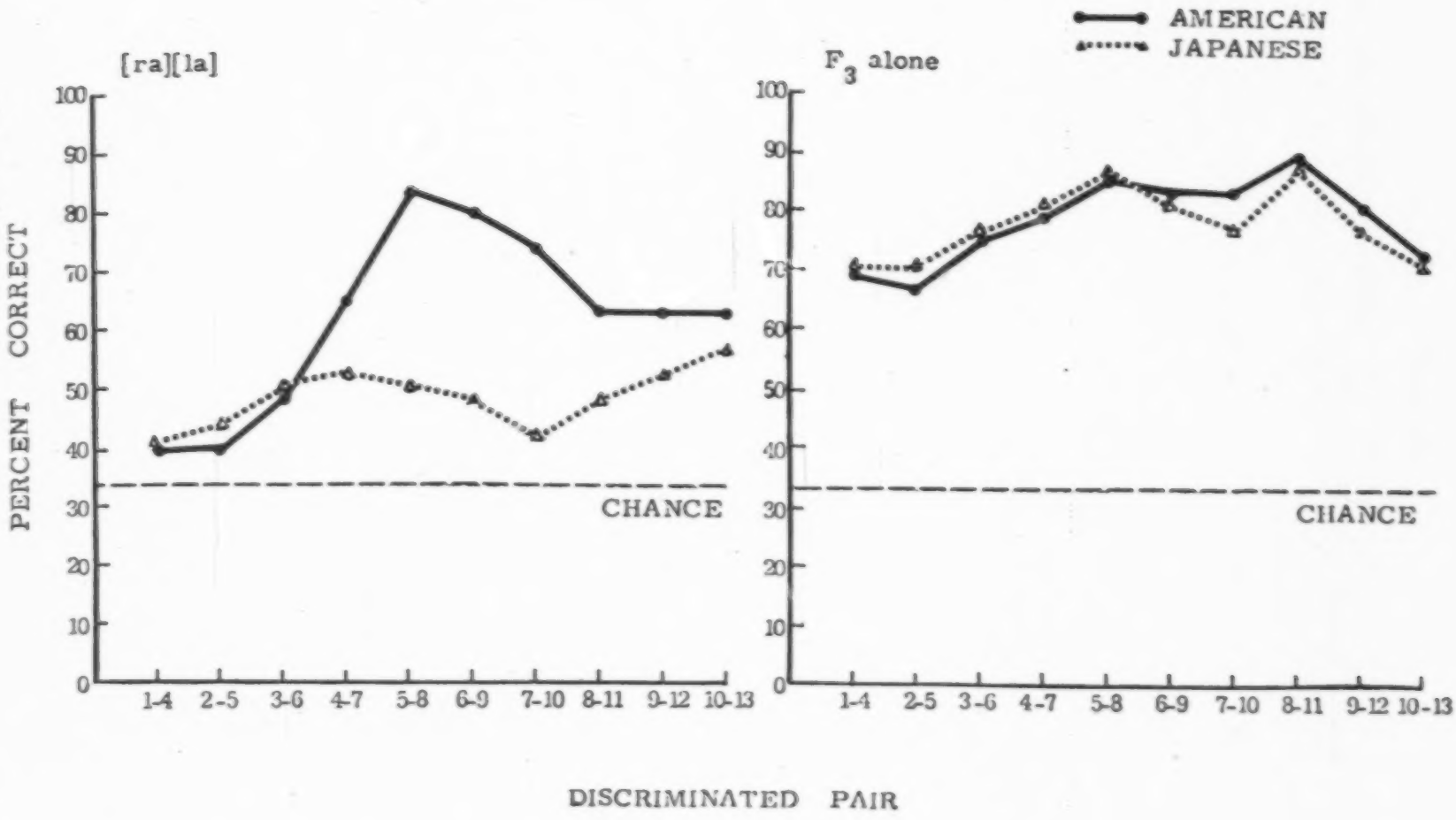


Figure 6: Overall American vs. overall Japanese.

REFERENCES

Abramson, A. S. and L. Lisker. (1970) Proceedings of the 6th International Congress of Phonetic Sciences, Prague.

Stevens, K. N., A. M. Liberman, M. Studdert-Kennedy, and S. E. G. Öhman. (1969) Lang. and Speech 12, 1-23.

Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results)

P. W. Nye and J. H. Gaitenby
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

In a preliminary study conducted at the University of Connecticut during 1971 and 1972, appraisals of the acceptability and intelligibility of texts generated in synthetic speech by rule were obtained from six blind students. The most critical of their comments expressed the view that the voice quality of the speech sounded excessively "nasal," and indicated that the words often missed or misheard were monosyllabic content words (in context). Other remarks indicated that a large part of the content of the speech was understood, but the study did not attempt to verify this objectively. A detailed examination of the intelligibility of synthetic speech from the phonetic level upward to the level of comprehension was clearly needed. Hence, late in 1972, plans were laid for tests to examine the lowest level of synthetic speech intelligibility using monosyllabic words as stimuli. This paper describes the results of the first of these intelligibility tests carried out on synthetic speech generated by rule. Later papers will describe the results of higher-level tests.

The objective of the current test was to isolate the consonants in synthetic speech that are less than optimally intelligible. However, to define an optimum level, a control condition was essential. Accordingly, it was decided that the subjects would also be exposed to a parallel test of the same stimuli in natural speech. Thus, in brief, the tests were to achieve the following:

a) To compare the intelligibility of synthetic speech with that of natural speech under quite severe test conditions (monosyllabic words in isolation).

b) To find the consonants, and the syllable location of those consonants, that are poorly perceived in synthetic speech, and thus to isolate areas requiring further research.

Acknowledgment: Research support for this work was provided by the Prosthetics and Sensory Aids Service of the Veterans Administration, with the assistance of a University of Connecticut Research Foundation Grant to Dr. J. David Hankins of the University. We thank Dr. Hankins for his assistance on a variety of problems. We also wish to express our appreciation to Nina deJongh, Margaret Allen, and Lee Donald for their administration of the test sessions and for tallying the data.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

c) To establish initial performance norms against which the results of future tests might be compared.

Early in the planning of the testing program the question arose as to whether a test should be specially designed or a so-called "standard" test should be used. The merits of the standard test are that they offer performance norms for natural speech against which the results of similarly obtained data may be compared and the techniques of administration thereby verified. During the last 30 years a considerable amount of developmental work has been carried out on a variety of intelligibility tests. Among the more well known are the Harvard PB lists of Egan (1948), the CID W-22 test (Hirsh, Davis, Silverman, Reynolds, Eldert, and Benson, 1952), the Fairbanks Rhyme Test (Fairbanks, 1958), and the CNC lists of Lehiste and Peterson (1959). From among the published results of these tests, two reports have emerged which claim that data from a modified version of the Fairbanks rhyme test show extremely small practice effects and are quite stable--even when obtained from inexperienced listeners (House, Williams, Hecker, and Kryter, 1965; Williams, Hecker, Stevens, and Woods, 1966). The absence of the need to train a panel of listeners gives the Modified Rhyme Test (MRT) a significant procedural advantage over many other tests. Accordingly the MRT was adopted for this study.

Modified Rhyme Test

The genealogy of the rhyme test may be traced back to its originator, Fairbanks (1958), who devised lists of words in which such factors as familiarity and phonetic frequency were taken into account. In the later version of House et al. (1965), which led to the present MRT, both of these restrictions were dropped and a closed-response set was instituted as well. The current word ensemble is shown in the typical response form illustrated in Table 1. Three hundred monosyllabic words are grouped on the form into 50 separate response sets of six words each; the words of each set occupy a numbered block and share certain phonetic properties. Nearly all of the words are of the consonant-vowel-consonant (CVC) type and, within each response set the six words share either an initial or final consonant. In 24 of the 50 sets, the shared consonant is in the initial position and in the remaining 26 sets the shared consonant takes the final position. The test consonants are inserted in the vacant space in each set. A full list of the test consonants and vowel environments employed in the MRT is given in the distributions shown in Table 2.

As a consequence of the rhyming condition and the fact that all test utterances are real English words, some significant gaps occur in the range of phonetic elements employed in the MRT. For example, the phonemes /ɹ̩/, /y/, /a/, /ə/, /u/, /ɝ/, /aI/, /aʊ/, /oʊ/, are not used at all, the phonemes /tʃ/ and /z/ are not found in initial position and the phoneme /ʃ/ does not appear in final position. These omissions should be kept in mind when reviewing the data. In addition, eight phonemes appear six times or less among the 300 words. These are /θ/, /ɚ/, /v/, /ʃ/, /ʒ/, /tʃ/, /dʒ/, and /ŋ/.

METHODS

Preparation of Test Materials

Test materials were generated in synthetic speech and in natural speech. The synthetic speech words were synthesized by rule (Mattingly, 1968) from a

TABLE 1: Response Sheet

MODIFIED RHYME TEST

NAME _____ DATE _____

1.		2.		3.		4.		5.	
game	came	sack	sag	dull	dub	not	hot	rave	rake
fame	name	sass	sat	duck	dun	got	tot	raze	ray
tame	same	sad	sap	dug	dud	lot	pot	race	rate

6.		7.		8.		9.		10.	
peel	keel	pen	den	back	bath	hip	dip	beat	seat
eel	reel	men	hen	ban	bat	sip	rip	heat	meat
heel	feel	then	ten	bad	bass	tip	lip	neat	feat

11.		12.		13.		14.		15.	
sun	sub	pale	pace	red	fed	dark	park	map	mass
sup	sung	pay	page	wed	shed	bark	mark	man	math
sum	sud	pane	pave	led	bed	hark	lark	mat	mad

16.		17.		18.		19.		20.	
boil	toil	save	sane	nun	gun	puff	put	teak	team
soil	foil	same	safe	sun	run	pun	pub	tear	tease
oil	coil	sake	sale	fun	bun	pug	pup	teal	teach

21.		22.		23.		24.		25.	
pill	pig	beat	beach	tang	tab	heat	heath	saw	raw
pick	pin	beak	bean	tam	tan	hear	heave	thaw	paw
pit	pip	beam	bead	tap	tack	heap	heal	law	jaw

26.		27.		28.		29.		30.	
din	pin	din	dip	went	rent	pick	tick	way	pay
tin	sin	did	dim	sent	dent	wick	sick	may	gay
win	fin	dill	dig	bent	tent	kick	lick	day	say

31.		32.		33.		34.		35.	
pale	sale	hill	bill	seed	seep	peak	peace	cup	cuff
bale	gale	will	fill	seethe	seem	peat	peach	cub	cud
tale	male	till	kill	seek	seen	peas	peal	cut	cuss

36.		37.		38.		39.		40.	
pig	rig	pang	pan	hang	sang	bun	bus	cold	hold
big	wig	pass	pad	bang	gang	buck	buff	gold	fold
fig	jig	path	pat	fang	rang	but	bug	sold	told

41.		42.		43.		44.		45.	
kid	kill	must	rust	cape	came	shop	mop	sit	sing
king	kit	gust	just	cake	cane	hop	top	sill	sin
kiss	kith	bust	dust	case	cave	pop	cop	sick	sip

46.		47.		48.		49.		50.	
rest	vest	hit	bit	fig	fin	shook	look	lame	lane
test	nest	wit	sit	fib	fit	took	cook	lay	lace
west	best	kit	fit	fizz	fill	book	hook	late	lake

TABLE 2

DISTRIBUTION OF TEST CONSONANTS AND VOWEL ENVIRONMENTS
USED IN THE MODIFIED RHYME TEST

	Consonants		Vowels (& V + n/r/l)	
	Initial Position	Final Position		
p	11	12	i	42
t	14	16	ɪ	66
k	9	15	ɛ	18
b	14	6	æ	36
d	7	11	ə	0
g	8	7	ɑ	12
f	12	4	ɔ	6
v	1	5	ʊ	6
θ	1	5	u	0
ð	1	1	ɚ	0
s	14	12	ʌ	42
z	0	4	ɝ	0
ʃ	3	0	ɜ˞	0
ʒ	0	0	eɪ	48
tʃ	0	3	aɪ	0
dʒ	3	1	aʊ	0
h	12	0	oʊ	0
m	7	9	ɔɪ	6
n	5	18	yu	0
ŋ	0	5	ɛn	6
r	10	2	oʊl	6
l	7	11	ɑr	6
w	9	0		
y	0	0		

typed phonetic input by means of the Haskins Laboratories DDP-224 computer system and parallel formant resonance synthesizer. The natural speech was provided by AA, a male linguist at the laboratories who used a high quality microphone and Ampex tape recorder. In addition to the 300 test words, the computer and speaker AA each produced a carrier phrase and the spoken numerals 1 to 50. These utterances were all pulse code modulated at an 8 kHz rate and stored on a disc of the DDP-224 computer. Six randomized series-lists, labeled A through F were generated, each series-list comprising six 50-word selections (one word from each block). In each series, each of the 300 test words appeared once. The series-lists were used to control the order in which the words (both synthetic and natural) were retrieved from the disc, converted from digital to analog form, and recorded on magnetic audio tape. Thus the series-lists A through F were used to generate two parallel sets of tapes--one set in synthetic speech and the other copy in natural speech. As a consequence of the sampling process the recorded natural speech and synthetic speech signals were low-pass filtered below 3.5 kHz.

Mode of Presentation

The test presentation procedure for both synthetic and natural speech versions consisted of a spoken number (indicating the block number from which the text word was about to be selected) and a carrier phrase that was followed, after a very brief pause, by the test word. Thus a typical presentation took the form "Number five. Please mark the word..pub" at a speaking rate of 130 words per minute. Four seconds then elapsed to allow the listeners time to mark their response sheets, before the next presentation commenced. One-hundred-fifty test words were presented at each session representing one-half of a series-list.

Tests were administered in a sound-damped booth with five subjects per session. Each subject wore a pair of Grason-Stadler binaural earphones type TDH39-300Z with ear muffs and heard the test words at a speaking level of approximately 80 db SPL. A partially balanced experimental design was adopted in which the groups of subjects heard natural speech tests and synthetic speech tests at alternating sessions. During each session the subjects completed three response forms of the type shown in Table 1. To avoid contextual effects arising from word order or spatial response biases for words within each block, both the word positions within groups and the group positions within blocks were permuted from test form to test form.

Experimental Subjects

Six groups of five undergraduate students from the University of Connecticut were employed as experimental subjects and paid for their time. The subjects were screened to eliminate those with defective hearing and were then given a practice session in synthetic speech using five blocks of words not included among the 300 words used in the actual test. The practice period was short and did not provide these naive listeners with any significant amount of prior training with synthetic speech. The purpose of the session was to familiarize the subjects with the presentation mode. The six groups of students encountered the speech materials in the order shown in Table 3. During the four-second interval between presentations the subjects were required, under forced choice conditions, to mark the word they had heard (or thought that they had heard) from among the six words appearing in the specified block. The subjects did not receive any information about the accuracy of their responses until after they had completed the

TABLE 3: Session Order

GROUP	SESSION											
	1	2	3	4	5	6	7	8	9	10	11	12
1	A1S	B1N	A2S	B2N	E1S	F1N	E2S	F2N	C1S	D1N	C2S	D2N
2	A1N	B1S	A2N	B2S	E1N	F1S	E2N	F2S	C1N	D1S	C2N	D2S
3	C1S	D1N	C2S	D2N	A1S	B1N	A2S	B2N	E1S	F1N	E2S	F2N
4	C1N	D1S	C2N	D2S	A1N	B1S	A2N	B2S	E1N	F1S	E2N	F2S
5	E1S	F1N	E2S	F2N	C1S	D1N	C2S	D2N	A1S	B1N	A2S	B2N
6	E1N	F1S	E2N	F2S	C1N	D1S	C2N	D2S	A1N	B1S	A2N	B2S

Each of the above three-character codes identifies a magnetic audio tape. The meaning of the characters is explained below:

The letters A - F symbolize the six series-lists, each of which contained 300 presentations.

Each series-list was presented in two parts; the identity of the part being indicated by the second symbol of the code.

Six groups of five subjects each heard all six series-lists in different orders. Three of the two-part lists were presented in synthetic speech (identified in the code by the letter S). The remaining lists were presented in natural speech (indicated by the letter N).

Sessions containing synthetic speech presentations were alternated with natural speech sessions.

entire experiment consisting of 12 sessions.

Data Retrieval

Errors being of primary interest, the data retrieval procedure was concerned with securing information about the circumstances in which they were made. Thus the response forms received from the subjects were examined and the errors identified and coded for computer processing. The record of each error consisted of a number indicating the phoneme actually presented, a second number of indicating the subject's response, a third number indicating the block in which the error occurred, and a fourth number indicating the test series. Using the computer record of the permuted ordering of the blocks in conjunction with a sorting routine, it was possible to create convenient tables of the errors made in each word group in each of the phonetic categories represented. The results are reported in the next section.

RESULTS

Learning Phenomena in Synthetic Speech

Both House et al. (1965) and Williams et al. (1966) have concluded that subjects employed in the MRT show very little evidence of learning. Their performance level with natural speech under a given set of conditions is uniform over a sequence of test sessions. This observation is essentially verified in the natural speech data obtained from the current test sessions. Figure 1 shows that, within a margin of $\pm$ 1.5 errors per session, the number of errors made on 150 natural speech presentations per session remains substantially constant over the entire test sequence. However, in successive exposures to synthetic speech, subjects show quite strong evidence of learning. Figure 1 also shows a graph of the average number of errors made by the 30 subjects during each of their six test sessions. A progressive improvement in performance with synthetic speech is evident throughout the sequence. If the subjects had been able to maintain continually the performance level achieved in their last synthetic speech session, their average error rate over the entire sequence would have fallen by more than 2.5 percentage points.

Overall Error Rates for Natural and Synthetic Monosyllables

The error rate for both syllable positions combined was 2.7% in natural speech and 7.6% in synthetic speech. The data for natural speech are consistent with the figure of 4% obtained monaurally by House et al. (1965) under their best signal-to-noise ratio of + 4 db, indicating that the correct procedure was applied in the present test. Closer inspection of the data shows that the natural speech stimuli elicited fewer errors in syllable initial position (1.9%) than in final position (3.5%). These figures reflect the well-known fact that the initial consonants of normal speech are more easily recognized than the final consonants. In contrast to the natural speech data, error rates in synthetic speech were nearly equal in the two positions: 8.0% initially and 7.2% finally. Recognition of the final consonants thus proved to be slightly better than recognition of initial consonants in synthetic speech.

Figure 2 shows the percentage of error for each phoneme that was confused

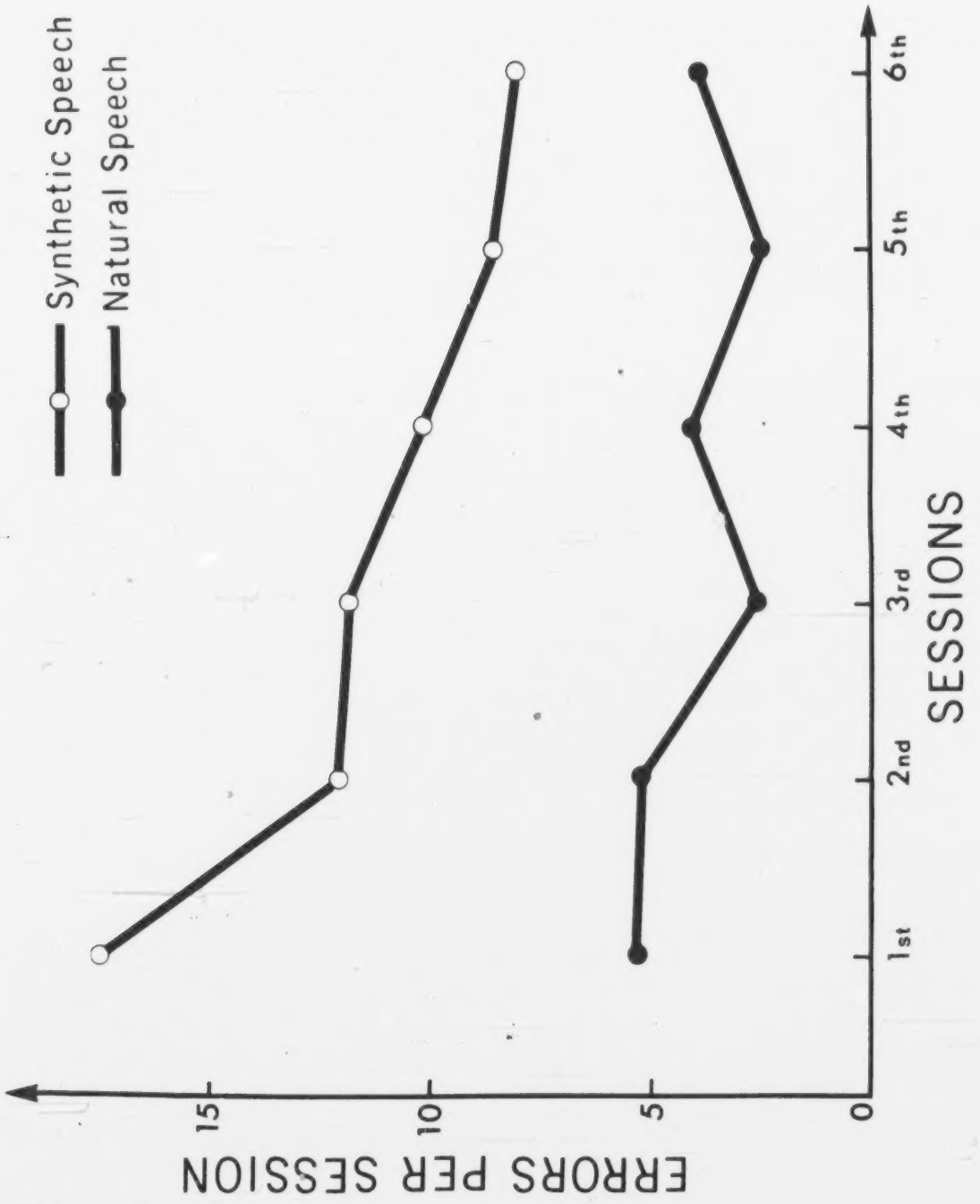
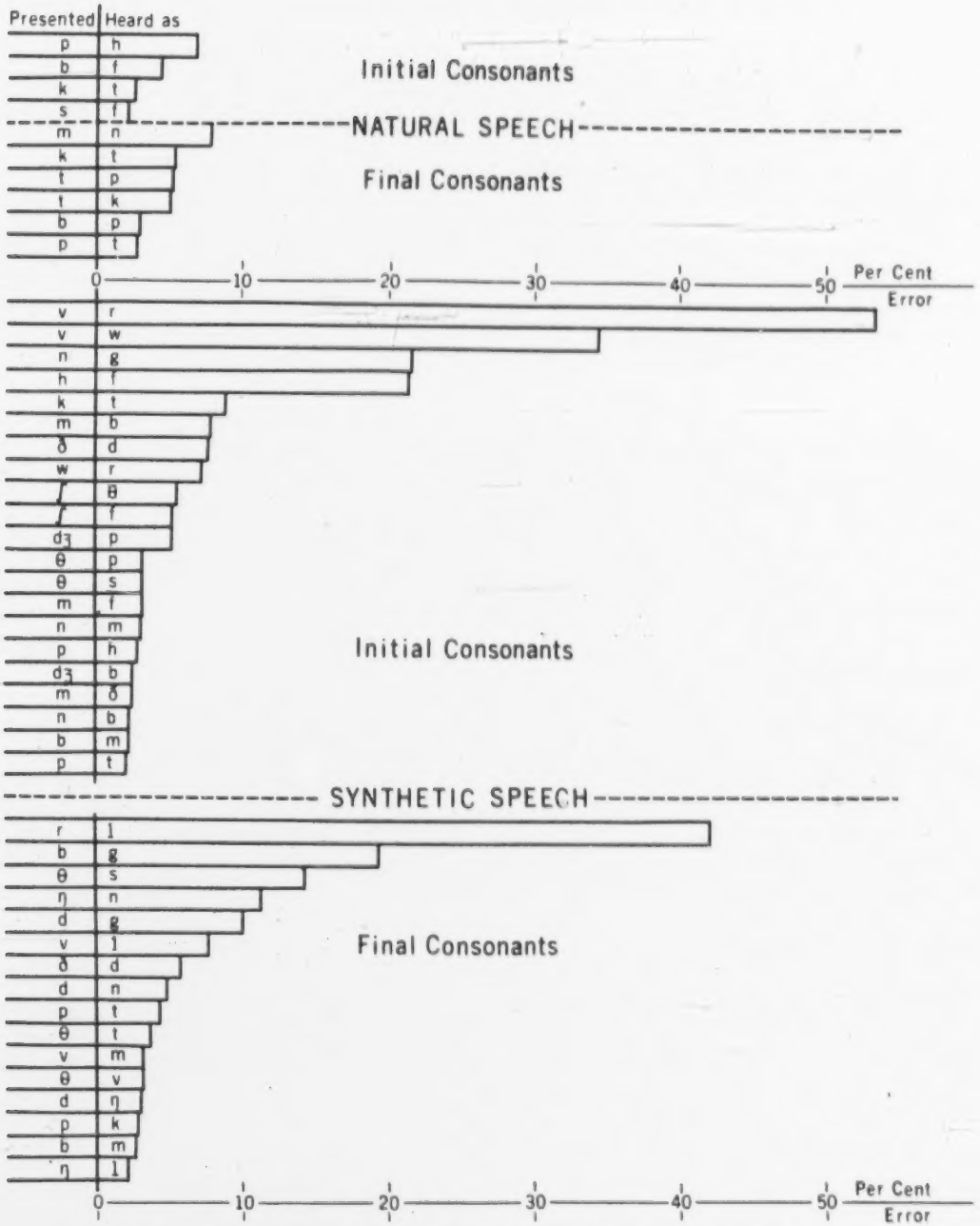

Synthetic Speech
Natural Speech
ERRORS PER SESSION
15
10
5
0
1st
2nd
3rd
4th
5th
6th
SESSIONS

Figure 1



PHONEMES RANKED BY PER CENT ERROR

Figure 2

in more than 2% of its appearances.[1] In synthetic speech, many intended phonemes were confused with more than one other phoneme. For example, in initial position, intended /v/ was heard as /r/ in some 50% of the presentations of /v/; and /v/ was heard as /w/ in some 30% of the intended /v/ presentations. Each of the two sounds confused with /v/ therefore appears in a separate row in Figure 2.

It must be reiterated that the MRT procedure requires the listener to choose one word from six in the printed set, as the word matching (or most closely matching) the one presented. If the listener fails to identify the actual correct word in the list, five alternative words remain. Unfortunately, these alternative words do not necessarily contain test consonants that are acoustico-phonetically similar to the consonant in the stimulus word, and the listener is thus forced to select a word that may not be what he heard at all, but only a word that is more or less like what he heard. A case in point is again the /v/ phoneme, which appeared in only one test word in initial position: the word "vest." Among the five alternative words there was none beginning with /f/ or /θ/ or /ð/--the most reasonable substitutions for /v/, phonetically and acoustically. Now if the assumption is made, and it is admittedly a considerable assumption, that the presented consonant is itself a reasonable acoustic facsimile of the correct intended consonant, it is evident that the confusions reported are of limited use in delineating specific problems in the rules for synthesis. The reader is accordingly cautioned to attend chiefly to the phoneme that produced the error, rather than to the phoneme(s) with which it was confused. This caveat applies in general to the error data--and obviates the necessity for undertaking more than a general survey of the particular confusions which were made.

Natural Speech Errors

In natural speech it was the voiceless phonemes, dominated by stops, that produced the largest number of confusions. Three of the four initial position phonemes with noticeable error, /p/, /b/, and /k/, also produced errors in final position. The main type of error indicated was Place,[2] although both /p/ and /b/ were heard as fricatives (Manner error) in initial position. Half of the total separate confusions involved labials, which is in line with classical confusion study results (Miller and Nicely, 1955).

Synthetic Speech Errors

Twelve phonemes produced significant error in initial position, but only eight in final position. Five phonemes produced confusions both initially and finally: /v/, /ð/, /θ/, /p/, and /b/. However, the amount of error for each

---

[1] An average error rate of 2% was considered to be an acceptable level on the basis that this figure was obtained in the natural speech data for initial position (the position with least error).

[2] The terminology of articulatory phonetics will be used to describe the various error types, e.g., Manner, Place etc.

differed greatly according to the syllable position of the phoneme (except for the phonemes /ð/ and /p/, which produced similar numbers of errors in both positions). Thus the phonetic classes confused in either position were the labiodental and dental fricatives (except /f/) and the labial stops. Voicing was the sole dimension rarely misheard.

Synthetic speech: initial position errors. The twelve consonants that were confused in initial position represented every Place and Manner of articulation. Both voiced and voiceless consonants produced errors, among which errors of Place dominated, although many errors involved simultaneous confusions in both Place and Manner. The intelligibility of /v/ was exceptionally poor in initial position. Also high in errors was /n/ which tended to be heard as a voiced velar stop, /g/. /h/ was often heard as another fricative, /f/, and /ʃ/ was heard as the fricative /θ/ or /f/ in a large number of cases. There were also a sufficient number of presentations (and variety of alternatives) of /m/ as well as /n/ to show that initial nasals were quite likely to be heard as stops. (The synthesizer employed to generate the stimuli lacked a nasal resonance channel and also had fricative constraints, so there is reason to suspect that Manner rules for synthesis may be less deficient than the list of errors might at first indicate.)

Errors of lesser degree occurred in /k/ (heard chiefly as /t/), and in /ð/ and /dʒ/ (both heard as stops). /w/ was heard as /r/ with some frequency. /θ/, /p/, and /b/ also produced errors, but /p/ and /b/ were, in fact, heard better in synthetic speech in initial position than they were in natural speech.

Synthetic speech: final position errors. Six of the eight phonemes that caused confusion when in final position were voiced consonants. It is noteworthy that /r/, which was the least intelligible in final position (post vowel), produced no significant error initially. In contrast, /b/, which produced relatively few confusions initially, was highly confused in final position. Like /d/ (which had no initial error), /b/ was heard chiefly as /g/ and occasionally as a nasal. /θ/, presented a number of times in final position, tended to be confused with /s/. /ŋ/ was often heard as /n/ (but was not heard as a stop, as was common with initial nasals). The intelligibility of /v/ was much greater finally than initially, but it still produced a large number of confusions. The fairly low error rates for /p/ (most often confused with /t/) and /ð/ (confused only with /d/) were very similar to those they produced in syllable-initial position.

Synthetic speech: the highly intelligible phonemes. Viewing the data from the positive side, there were four phonemes that were clearly intelligible in either syllable position: /t/, /g/, /f/, and /s/. Phonemes with good initial-position intelligibility (but not final) were /r/ and /d/; those that were good in final position (but not initially) were /n/, /m/, /k/, /d/ and two that were presented only in final position, /z/ and /tʃ/. When arranged in a conventional phonetic chart form, these positive data appear as in Figure 3.

Figure 3 shows that as a group the alveolar phonemes were the best perceived. The second formant (F_2) target frequency required by alveolar sounds is intermediate between the frequencies of the highest and lowest F_2 vowel formants, and thus relatively small adjustments in F_2 transitions are required for normal movement from alveolars to any adjacent vowels (or consonants). On the

Figure 3

SYNTHETIC SPEECH: PHONEMES WITH HIGHEST INTELLIGIBILITY

	LABIAL	LAB-DENT	DENTAL	ALVEOLAR	PALATAL	VELAR	GLOTTAL
Voiceless STOP				t		k[2]	
Voiced STOP				d[1]		g	
Voiceless FRICATIVE		f		s	tʃ[*2]		
Voiced FRICATIVE				z[*2]	dʒ[2]		
NASAL	m[2]			n[2]			
LIQUID				r[1]			

LEGEND

Phoneme presented in final position only: *

Phoneme good in initial position only: 1

Phoneme good in final position only: 2

(/ʒ/ and /y/ did not appear in the test.)

other hand, the least intelligible consonants (the labials and labiodentals) require F_2 transitions from very low frequencies and thus large transition adjustments must occur for normal coarticulation with the upper range of vowels. It should be noted that most of the vowel environments appearing in the test syllables were high front vowels, i.e., those with F_2 centers at the most extreme distance from the labial F_2 targets.

CONCLUSIONS

Summary

The MRT has been widely used as a test of intelligibility in Voice Transmission Systems and to assess hearing loss (Williams et al., 1966; Kreul, Nixon, Kryter, Bell, Lang, and Schubert, 1968). It may thus be considered to qualify as a standard test and hence a reasonable test to use--certainly as a point of departure--for synthetic speech evaluation. The first purpose of the study was to compare the intelligibility of monosyllables in synthetic speech and natural speech and to establish an acceptable intelligibility level. For that task the MRT was adequate, although not entirely ideal. The results of the test showed an overall intelligibility score for monosyllabic synthetic speech of 92.4% and a score of 97.3% for natural speech. Future work on the rules for synthesis will be directed toward closing this 5% gap.

The second purpose of the test was to identify those consonants that produce significant recognition errors (confusions). The results are clear in pointing out the consonants that are confused. The least intelligible of the synthesized phonemes were initial /v/ and final /r/ (vowel plus /r/). The general types of phonemes shown to be most in need of improvement in the rules are the labials, labiodentals, and dentals--in both initial and final positions. The Manner least well synthesized overall was frication (in both positions). There were also numerous errors in the nasalization Manner (nasals heard as stops, and stops heard as nasals--in both positions) indicating (within the test limitations) that additional work must be done in this area, although nasal confusions probably result as much from the synthesizer employed (which had no nasal channel) as from the rules for synthesis of the sounds.[3] This point in fact applies quite generally to all the error data. When synthetic speech is tested there are three major variables involved: 1) the rules of synthesis (phonemes, allophones, prosodic features); 2) the synthesizer used to generate the speech; and 3) the "power" of the testing procedure used to distinguish the intelligible from the unintelligible consonants and vowels. It will therefore be clear that our ability to refine and improve synthetic speech depends not only on selecting appropriate tests but also on ensuring that the results of these tests are actually used to improve the rules and to forward the development of synthesis equipment that is better adapted to reading machine requirements.

A third purpose of the MRT was to provide an opportunity to compare synthetic speech with natural speech and with the performance norms obtained with the

---

[3] Another synthesizer, with nasal resonance circuitry and an improved frication facility, will be available for the preparation of subsequent tests.

natural speech used by other workers. Broadly, the overall results of the tests agreed well with the data obtained by House et al. (1966) and Kreul et al. (1968) on normal listeners, and indicated that the speaker AA did not over-articulate and that the PCM encoding process used to construct the tapes did not introduce any serious distortions. However, when examined in close detail, the data obtained from the synthetic speech tests emphasize various weaknesses in the MRT design.

Limitations of the MRT

For example, the MRT suffers from an imbalance in the number of times various consonants are presented, an uneven representation of initial and final consonants (with two phonemes omitted entirely), and an incomplete representation of vowel environments. These negative features represent shortcomings even in a natural speech test but when applied to synthetic speech they prove to be particularly serious because several phonemes that were poorly discriminated occurred very infrequently in the test. This is important in light of the fact that our study has shown that there is a significant learning process involved in the use of synthetic speech which can be traced throughout the test. Therefore, we cannot overlook the possibility that the relatively low frequency of occurrence of some phones gave scant opportunity for learning and may have contributed to the low intelligibility scores for those phones.

Future Plans

In formulating a strategy for future tests of the intelligibility of synthetic speech, it is clear that we must next explicitly identify the confusions that occur among the poorly perceived phonemes isolated in the current test. It is equally clear that as a practical matter, the available stimuli for use in studying these phonetic confusions are either nonsense syllables or real words, while the response modes can be either open or closed. (The MRT used real words in a closed response mode.)

The ideal test for identification of specific kinds of errors made would employ nonsense syllables as stimuli, and trained listeners reporting the sounds heard in phonetic symbols (open response). However, there are many practical difficulties which hinder the recruiting of trained listeners and we are obliged to continue the evaluation study with relatively unskilled listeners. This, in effect, reduces the options in test methodology to two.

The first option then, acknowledging the fact that phonetically naive listeners can report only through the medium of standard orthography, involves the use of real words in an open response formant. One version of this type of test, words in sentence-like strings--grammatically correct but meaningless--is already underway.

The other remaining option uses nonsense syllables in a closed response situation. For maximum effectiveness, such a closed response test would be designed to use sets of alternative consonants which are phonetically much closer to the presented stimuli than is the case in many of the MRT word sets. In such a test special attention must also be made to ensure that the vowel environments used include more back vowels and diphthongs than are present in the MRT.

Both of these types of test represent practical compromises with the ideal test, but it is anticipated that they will nevertheless reveal specific deficiencies in the rules for synthesis of initial and final consonants without overtaxing the willing but unskilled listeners.

REFERENCES

Egan, J. P. (1948) Articulation testing methods. Laryngoscope 58, 955-991.

Fairbanks, G. (1958) Test of phonemic differentiation: the Rhyme Test. J. Acoust. Soc. Amer. 30, 596-600.

Hirsh, I. J., H. Davis, S. R. Silverman, E. G. Reynolds, E. Eldert, and R. W. Benson. (1952) Development of materials for speech audiometry. J. Speech Hearing Dis. 17, 321-337.

House, A. S., C. E. Williams, M. H. L. Hecker, and K. D. Kryter. (1965) Articulation-testing methods: consonantal differentiation with a closed-response set. J. Acoust. Soc. Amer. 37, 158-166.

Kreul, E. J., J. C. Nixon, K. D. Kryter, D. W. Bell, J. S. Lang, and E. D. Schubert. (1968) A proposed clinical test of speech discrimination. J. Speech Hearing Res. 11, 536-548.

Lehiste, I. and G. E. Peterson. (1959) Linguistic considerations in the study of speech intelligibility. J. Acoust. Soc. Amer. 31, 280-286.

Mattingly, I. G. (1968) Synthesis by rule of General American English. Ph.D. dissertation, Yale University. (Issued as Supplement to Haskins Laboratories Status Report on Speech Research.)

Miller, G. A. and P. E. Nicely. (1955) An analysis of perceptual confusions among some English consonants. J. Acoust. Soc. Amer. 27, 338-352.

Williams, C. E., M. H. L. Hecker, K. N. Stevens, and B. Woods. (1966) Intelligibility test methods and procedures for evaluation of speech communication systems. National Technical Information Service AD 646-781.

Forward and Backward Masking of Brief Vowels*

M. Dorman,[+] D. Kewley-Port, S. Brady-Wood,[++] and M. T. Turvey[++]
Haskins Laboratories, New Haven, Conn.

From an information processing viewpoint (Haber, 1969a), perception can be viewed as a hierarchically organized set of operations which extract over time different categories of information from a signal. In studies of visual perception, backward masking of form by a patterned stimulus has become a powerful technique for probing stages of perceptual processing (Haber, 1969b; Turvey, 1973). The use of this technique is based on the assumption that when a masking stimulus follows a target stimulus after some delay, processing of the target stimulus occurs during the delay, but further processing is either distorted or interrupted by the arrival of the mask. By varying the interval between target and mask, successive operations in the extraction of stimulus information can be probed.

Although peripheral and central operations in vision have been isolated and investigated by systematic variation of the physical and temporal parameters of target and mask stimuli in forward and backward masking paradigms (Turvey, 1973), few studies have used these techniques to probe the recognition of speech. Massaro (1972) presented listeners 20 msec vowels, either /i/ or /I/, followed at interstimulus intervals (ISIs) from 0 to 500 msec by a 270 msec nonsteady-state mask. Recognition of the target vowels was near chance at 0 msec ISI and reached asymptote by 250 msec ISI. Pisoni (1972) presented listeners 40 msec vowels, /i/, /I/, or /ɛ/, followed at ISIs of 0 to 450 msec by another vowel from the same set. Target vowel recognition was 80% correct at 0 msec ISI and reached asymptote at 80 msec ISI. While these studies serve as a useful beginning, Turvey (1973) has pointed out that inferences about perceptual processing from studies employing masking techniques can best be made only after systematic variation of the relationship between target and mask stimuli. This paper reports the first two of a series of experiments which used forward and backward masking paradigms to investigate perceptual operations in the recognition of brief vowels. Experiment I determined the vowel duration and ISI necessary to evade forward and backward masking. Experiment II then compared the masking produced by the computer synthesized vowel-like stimulus of Experiment I, this same stimulus 20 db more intense, and a real-speech vowel.

---

*Revised version of a paper presented at the 85th meeting of the Acoustical Society of America, Boston, Mass., April 1973.

[+]Also Herbert Lehman College of the City University of New York.

[++]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

EXPERIMENT I

To determine the vowel duration necessary to evade interference from either a preceding or following stimulus, vowel sets of three durations [15.5, 20, and 30 msec] were constructed. In order to make the mask distinctive from the target vowels, a two-formant vowel-like mask was synthesized. To maximize the probability that the mask would be effective, the mask duration was 125 msec.

Method

Subjects. The Ss were undergraduates from Yale University and the University of Connecticut. Yale University students received $2.00 per hour for participation. University of Connecticut students received class credit.

Apparatus. The stimuli were recorded and reproduced on an Ampex AG500 tape recorder. The tape recorder output was interfaced with a distribution amplifier which insured equal signal amplification into four sets of matched Grason-Stadler TDH39-300Z earphones. The stimuli were presented binaurally at a comfortable listening level. A calibration signal insured equal signal levels in all conditions within an experiment.

Preparation of stimuli. The target stimuli were the vowels /i/, /ɛ/, /ʌ/ spoken by a male voice with a fundamental frequency of approximately 120 Hz. With the aid of the Haskins Laboratories computer controlled PCM system three sets of vowels were prepared. For the sets the vowels were truncated to durations of 15.5 msec, 20 msec, and 30 msec. The mask was a computer synthesized two-formant sound of 125 msec duration with formant frequencies at 489 Hz and 1690 Hz. This mask was vowel-like but did not have formant frequencies similar to any English vowel. The target and mask stimuli were equated for peak to peak amplitude.

Training materials. Under computer control one sequence of three repetitions of the vowel set /i ɛ ʌ/ and six 18-item sequences of target vowels (six repetitions of each vowel in each randomized sequence) were recorded on audio tape. The intertrial interval was 4 sec for all sequences.

Test materials. Under computer control, six test sequences were constructed. For each vowel set duration [15.5, 20, 30 msec] both a forward and backward masking sequence were generated. For each test sequence a six-item practice sequence was also generated. In the forward masking condition the mask preceded the target vowels at intervals of 0, 25, 50, 100, 200, and 500 msec. Each vowel occurred six times at each ISI. In the backward masking condition the vowels preceded the mask at intervals of 0, 25, 50, 100, 200, and 500 msec. Each vowel occurred six times at each ISI. The sequence of vowels and ISIs was randomized in each test sequence. Each test sequence was presented twice, thus creating two blocks of 54 trials, or a test sequence of 108 items.

Design. One group of Ss was trained and tested with 15.5 msec vowels, another group with 20 msec vowels, and a third group with 30 msec vowels. All Ss were tested on both the forward and backward masking sequences in counterbalanced order.

Procedure

The Ss were seated in a large sound attenuated room and were told they would hear three very brief vowels /i/, /ɛ/, /ʌ/ which they were to learn to identify. First, the Ss were presented three repetitions of the three-vowel set. Next, the Ss were told they would hear five 18-item lists of the vowels in random order and were instructed to write the identity of the vowels on printed response sheets. The correct responses, initially covered by a movable slider, were printed next to the space for the Ss' responses. By moving the slider down the page for each succeeding trial the Ss uncovered the correct responses for the preceding trial, thus providing immediate feedback of correct responses. Finally, the Ss were presented an 18-item test sequence with no feedback of correct responses.

After a brief rest period, the Ss were told they would hear, in one sequence, the vowels followed by a mask at various intervals, and in another sequence, the mask followed by the vowels at various intervals. The Ss were instructed to write the identity of the vowels on a printed answer sheet. After six practice trials the Ss were presented a 108-item test sequence in two blocks of 54 trials. Then, after a brief rest, the Ss were given another six practice trials and the other 108-item test sequence.

Results

Only those Ss who made no errors on the final (no feedback) practice sequence were considered in the data analyses. Post hoc inspection of errors as a function of ISI in the backward masking condition revealed two distinct error patterns. One population was characterized by better than 80% correct responses at the 0 msec ISI and very few errors at other ISIs. These Ss will be referred to as Nonmaskers. Another population was characterized by scores of less than 80% correct at 0 msec ISI while not reaching asymptotic performance until 100-200 msec ISI. These Ss will be referred to as Maskers.

The results for the 15.5, 20, and 30 msec vowel duration groups in the backward masking condition are shown in Figure 1. All 10 Ss trained with the 30 msec vowels achieved perfect performance on the final practice sequence. None of these Ss was characterized as a Masker on the test sequences. Of the Ss trained with the 20 msec vowels, 83% achieved perfect performance on the final practice sequence. Seven of the ten subjects in the group were classified as Nonmaskers, and three as Maskers on the test sequences. The Maskers did not reach asymptotic performance until 200 msec ISI. Of the Ss trained on the 15.5 msec vowels, 62% achieved perfect performance on the final practice sequence. Of the ten Ss in this group, six were classified as Nonmaskers, and four as Maskers on the test sequence. The Maskers' scores did not reach asymptote until 200 msec ISI. (The depressed score at 50 msec ISI reflected the poor performance of only one S.)

In the forward masking condition, all of the groups achieved 97% or better correct responses at all of the ISIs.

Discussion

No forward masking was observed in any of the conditions. Since the 15.5 msec vowels were of minimum duration (little more than one pitch period), it appears that perceptual interference in the recognition of vowels occurs only when a masking stimulus follows a target stimulus.

Figure 1

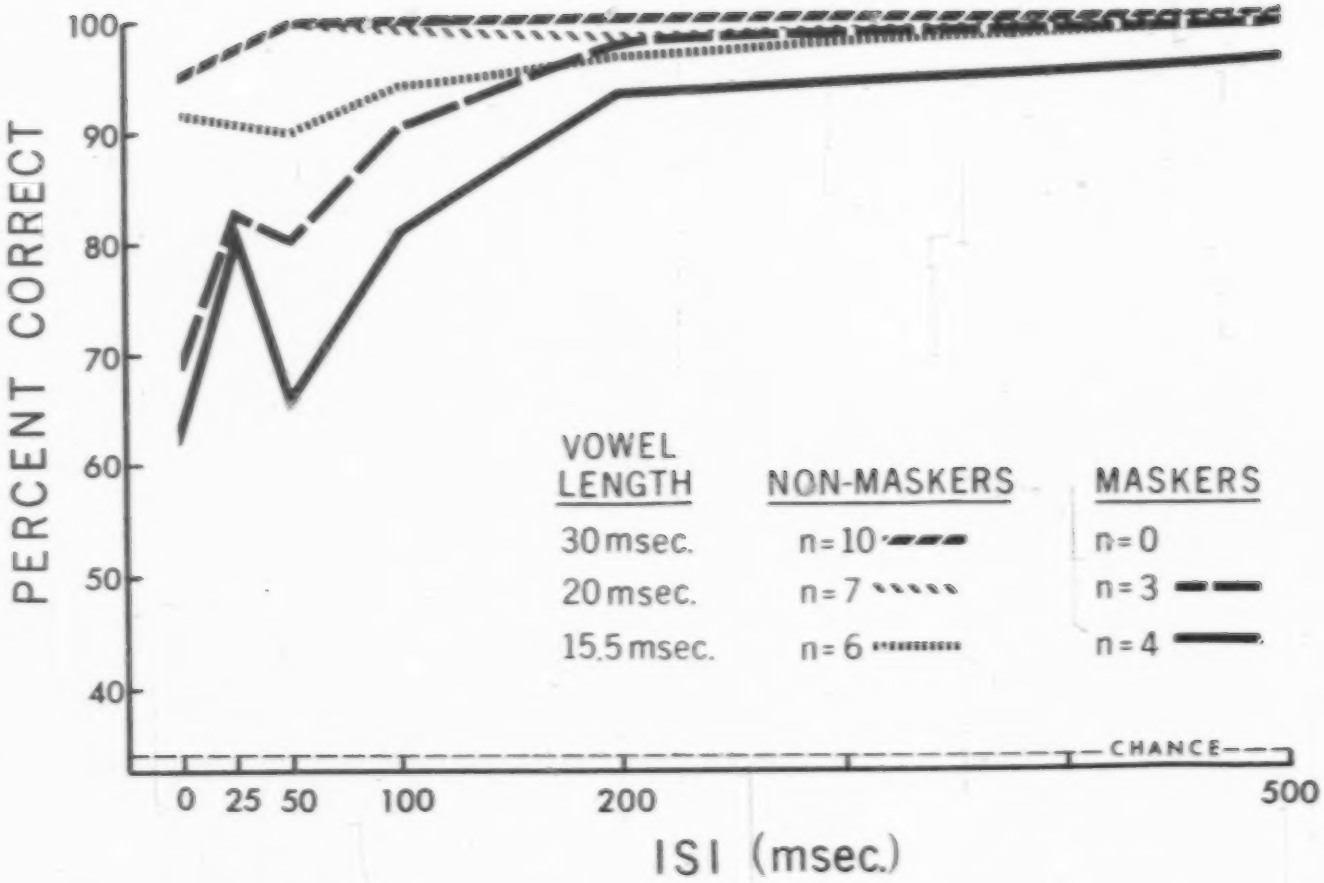


Figure 1: Average recognition scores for Nonmaskers and Maskers as a function of vowel length and interstimulus interval.

As was reviewed earlier, the assumption underlying the use of backward masking to study perceptual processing is that when a mask follows a target after some delay, processing of the target occurs during the delay, but further processing is interrupted by the arrival of the mask. The interpretation of the backward masking data which follows is based on this point of view.

In the 30 msec vowel condition at 0 msec ISI, recognition of vowel targets was essentially perfect. In the 20 msec vowel condition at 0 msec ISI, the majority of Ss showed very little or no impairment in vowel target recognition. Even when the vowels contained only one pitch period, and therefore had minimally defined vowel formant frequency location, the majority of Ss performed at better than 90% accuracy at 0 msec ISI. From these data we conclude that for the majority of Ss, a stimulus duration of between 20 and 30 msec is sufficient for processing mechanisms to separate the target vowels from the mask and to extract the features necessary for the recognition of the vowels. Of course, this conclusion applies only to the specific conditions of Experiment I (i.e., a three-vowel target set and a 125 msec vowel-like mask).[1]

It is important to note that the near perfect vowel recognition at 0 msec ISI, for at least the majority of Ss with the 15.5 msec vowels, was not simply a function of an "easy" discrimination (cf. Massaro, 1972:134). Only 62% of the Ss trained with the 15.5 msec vowels could identify the vowels in isolation after 90 practice trials. To determine whether Ss who could not discriminate the vowels in isolation would evidence a masking function similar to that reported by Massaro (1973), eight Ss who made errors (average = 79% correct) on the final training test with 15.5 msec vowels were tested in the backward and forward masking sequences. For these Ss vowel recognition did not markedly improve with an increase in ISI. At 0 msec ISI vowel recognition was 55% correct; at 500 msec ISI, 66% correct.

While the majority of Ss achieved near perfect vowel recognition at all ISIs, 30% of the Ss in the 15.5 and 20 msec vowel conditions were characterized as Maskers, i.e., vowel recognition at 0 msec ISI was approximately 60% correct and did not reach asymptote until 100-200 msec ISI. The long ISI necessary to evade masking for these Ss is in marked contrast to the essentially perfect performance of all Ss at 0 msec ISI in the 30 msec vowel condition. The magnitude of this difference suggests that those Ss characterized as Maskers may have employed rather different recognition routines to identify the vowel targets than the Nonmaskers. However, until other combinations of target and mask parameters have been manipulated, speculation on this point would be premature.

EXPERIMENT II

Since the mask of Experiment I was a synthesized two-formant vowel-like stimulus, it is possible that a mask which shared more features with the target vowels would produce more interference with vowel recognition. To investigate

---

[1] To determine whether this outcome was a function of the particular stimuli used in Experiment I, the 15.5 msec and 20 msec vowel conditions were replicated with the vowels /i, ɛ, ʌ/ spoken by a female informant. The same pattern of results were obtained.

this, in one condition of Experiment II the mask was a 125 msec vowel. To determine whether mask energy affects vowel recognition, in another condition of Experiment II the mask was the two-formant mask of Experiment I increased in energy 20 db.

Method

Subjects. The Ss were undergraduates from Yale University and the University of Connecticut. Yale University students received $2.00 per hour for participation. University of Connecticut students received class credit.

Preparation of stimuli. The target vowels were the 20 msec vowels used in Experiment I. Under computer control, two masks were constructed. One mask was the vowel /o/ spoken by the same male speaker as in Experiment I, truncated to 125 msec duration and equated for peak to peak amplitude with the target vowels. A second mask was the two-formant mask of Experiment I, but made 20 db (true RMS) more intense than the mask of Experiment I.

Test materials. Under computer control backward and forward masking sequences were generated with both the /o/ mask and the +20 db two-formant mask. The internal construction of the test sequences was the same as in Experiment I.

Design. One group of Ss was tested with the /o/ mask in both the forward and backward masking conditions in counterbalanced order. Another group of Ss was tested with the +20 db mask in the forward and backward masking conditions in counterbalanced order.

The Apparatus, Training materials, and Procedure were the same as in Experiment I.

Results

Only those Ss who made no errors on the final practice sequence were considered in the data analyses. The results for the Maskers in the /o/ mask and +20 db mask conditions plus the Maskers from the 20 msec vowel condition of Experiment I are shown in Figure 2. Of the Ss trained for the /o/ mask condition, 83% achieved perfect performance on the final practice sequence. Five of the ten Ss in this group were characterized as Nonmaskers, and five as Maskers on the test sequence. The Maskers did not reach asymptotic performance until 100 msec ISI. Of the Ss trained for the +20 db mask condition, 90% achieved perfect performance on the final practice sequence. Of these 10 Ss, eight were characterized as Nonmaskers, and two as Maskers on the test sequence. The Maskers did not reach asymptotic performance until 100 msec ISI.

In the forward masking conditions, both groups of Ss achieved 92% or better correct responses at all of the ISIs.

DISCUSSION

The absence of forward masking in either the /o/ mask or +20 db mask condition reinforces the impression gained from Experiment I that perceptual interference occurs only when a mask follows a target stimulus. The implication of this outcome is discussed below.

Figure 2

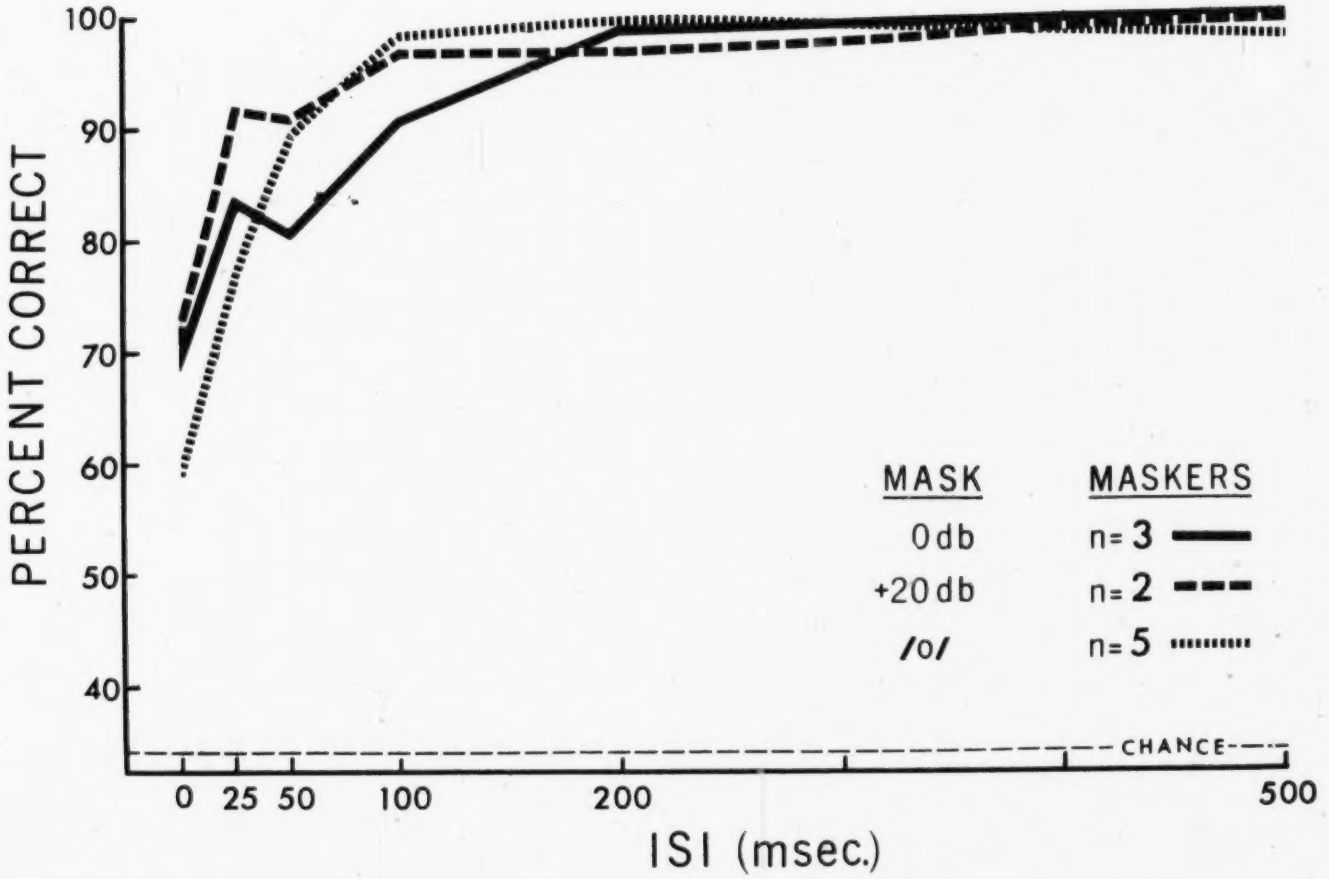


Figure 2: Average recognition scores for Maskers as a function of type of mask and interstimulus interval.

In the backward masking sequences the performance of the Ss did not differ greatly from that of the Ss in the 20 msec vowel two-formant mask condition of Experiment I. One group of Ss was characterized as Nonmaskers and a smaller group as Maskers. Increasing mask energy did not increase the number of Maskers or increase the ISI necessary to evade masking. The /o/ mask appeared to be somewhat more effective than either of the two-formant masks in terms of the number of Maskers, and percent correct vowel recognition at 0 msec ISI. Overall, however, the differences between the groups were small. Viewed as a whole, the results from Experiments I and II, i.e., the complete absence of forward masking and the absence of an increase in backward masking with a large increase in mask energy, suggest that the locus of interference for the Maskers was of central rather than peripheral or sensory origin (cf. Turvey, 1973:36). The data do not reveal, however, the nature of the difference in perceptual processing for the Nonmasker and Masker populations, or the mechanism responsible for the masking.

The outcome for Experiments I and II are in marked contrast to the data reported by Massaro (1972) and Pisoni (1972). Massaro reported near chance recognition of targets from a two-vowel set at 0 msec ISI. In the present study, even in the 15.5 msec vowel condition with a three-vowel set, the majority of Ss showed better than 90% target recognition at 0 msec ISI, while no S was near chance performance. The explanation for this difference in outcome is not readily apparent. Pisoni reported, for 40 msec vowels and a 40 msec mask, 80% target recognition at 0 msec ISI. In the present study all of the Ss in the 30 mse vowel condition at 0 msec ISI could identify the target vowels with essentially perfect accuracy. These data suggest that a stimulus which is equal in duration to a target stimulus may be a more effective mask than a stimulus which is much longer and more intense. This possibility is currently under investigation.

REFERENCES

Haber, R. N. (1969a) Information processing analyses of visual perception: an introduction. In Information Processing Approaches to Visual Perception, ed. by R. N. Haber. (New York: Holt, Rinehart, and Winston).

Haber, R. N. (1969b) Repetition, visual persistence, visual noise and information processing. In Information Processing in the Nervous System, ed. by K. N. Leibovic. (New York: Springer-Verlag).

Massaro, D. W. (1972) Preperceptual images, processing time, and perceptual units in auditory perception. Psychol. Rev. 79, 124-145(a).

Pisoni, D. (1972) Perceptual processing time for consonants and vowels. Haskins Laboratories Status Report on Speech Research SR-31/32, 83-92.

Turvey, M. T. (1973) On peripheral and central processes in vision: inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.

Effects of Proactive Interference and Rehearsal on the Primary and Secondary Components of Short-Term Retention

M. T. Turvey[+] and Robert A. Weeks[++]

Under the conditions of the distractor paradigm, short-term retention declines to a minimum in a very brief period. The rapid forgetting can be said to reflect the declining contribution of the short-term store or primary memory and the asymptote can be taken as a measure of the contribution of the long-term store or secondary memory. It was shown that manipulating proactive effects by varying the recency or the rehearsal of prior material affected only the primary memory component of the short-term retention function. On the other hand, manipulating the difficulty of the subsidiary task performed during the retention period with proactive effects held constant affected both the primary and the secondary components. The results were discussed with respect to the relation between the two memory components and the idea that proactive effects are limited to long-term store.

Since the renewal of interest in short-term retention (Brown, 1958; Peterson and Peterson, 1959) [the earlier work of W. Smith (1895) and T. Smith (1896) having been neglected] there have been two general views of postcategorical memory. According to one view, expressed elegantly by Melton (1963), the retention of material over both brief and long periods of time is subserved by the same mechanism. This view sought support in the demonstration that variables which affect long-term retention similarly affect short-term retention. Thus Melton (1963) in his defense of the unitary view of memory emphasized the experiments of Keppel and Underwood (1962) which had shown the susceptibility of short-term retention to proactive interference effects, and the apparent continuity of the effects of repetition upon retention.

Contrary to the unitary memory view is the idea that there are two quite distinct memory systems underlying the short-term and long-term retention of material--systems which probably obey different constraints and probably work with

---

[+]Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

[++]University of Connecticut, Storrs.

Acknowledgment: This research was supported in part by a grant from the University of Connecticut Research Foundation. The authors would like to acknowledge the assistance of G. Magalnick in the running of the experiments.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

different codes. This view was underscored in the writings of Hebb (1949) and Broadbent (1958) and has in more recent times been championed by Waugh and Norman (1965), Peterson (1966a), and Atkinson and Shiffrin (1968), among others. While arguing for two distinct memory systems recent advocates of this position generally agree that under some circumstances an item can be represented in both systems concurrently. The argument is made (e.g., Waugh and Norman, 1965; Peterson, 1966a) that recall in short-term memory situations probably reflects both the hypothesized short-term store or primary memory, and the hypothesized long-term store or secondary memory. Thus, while recall at longer intervals is determined solely by long-term store, at brief intervals both systems contribute, with the relative contribution of short-term store decreasing and the relative contribution of long-term store increasing as a function of time in memory (Peterson, 1966a) or of number of items (Waugh and Norman, 1965).

Currently the idea of two memory systems holds sway among students of memory. Accordingly, the present paper is concerned with identifying the effects of certain variables on the components of short-term memory performance. More specifically, the present paper asks whether certain sources of short-term forgetting, proactive interference and interpolated task difficulty, are exerting their influences in short-term store, long-term store, or both.

The experiments reported here made use of the short-term memory distractor paradigm (Brown, 1958; Peterson and Peterson, 1959). The forgetting functions obtained with this procedure can be analyzed along the lines suggested by two-process theorists. In general, these functions have been characterized by a rapid decline in recall which typically attains a low asymptote within 10 seconds (Baddeley and Scott, 1971; Dillon and Reid, 1969; Merikle, 1968; Posner and Konick, 1966; Schierer and Voss, 1969; Turvey, Brick, and Osborn, 1970b). The asymptotic component can be viewed as reflecting primarily long-term store strength and the retention loss prior to asymptote can be taken as reflecting primarily the rapidly declining contribution of short-term store (cf. Kintsch, Crothers, and Jorgensen, 1971; Peterson, 1966a). One of the major sources of forgetting in the short-term memory distractor situation is proactive interference and in the view of some authors (Craik and Birtwistle, 1971) the locus of proactive interference effects is in long-term store. Experiments II and III of the present series sought to localize more accurately the influence of proactive interference on short-term retention by varying the interval between successive short-term memory tests. The degree of short-term forgetting has been shown to be related inversely to intertest interval (e.g., Loess and Waugh, 1967).

Another major and obvious source of forgetting in the short-term memory distractor situation is the distractor, that is, the interpolated activity, and the degree to which it can be said to be antagonistic to rehearsal. Experiment IV looks at the effects of varying the degree of difficulty of the interpolated task on the two components of the short-term memory function. In short, Experiment IV seeks to determine the locus of the effect of preventing rehearsal. Experiment V, on the other hand, manipulates a prior item's potential for interfering by varying its opportunity to be rehearsed and then asks a question similar to that raised by Experiments II and III: where is the effect of this proactive interference manipulation registered in the short-term forgetting function?

With successive short-term memory tests of the distractor kind retention declines precipitously across tests, reflecting the development of proactive

interference (cf. Keppel and Underwood, 1962). But since the largest decrease in performance is generally from Test 1 to Test 2 in a series, it can be inferred that most proactive interference accrues with one prior test. A procedure using two short-term memory tests, therefore, affords the simplest technique for manipulating and examining proactive interference effects. In the present experiments subjects received a number of trials, each trial consisting of two successive short-term memory tests. The interval between trials was of sufficient length to prevent or at least minimize proactive interference effects carrying over from trial to trial. As a number of investigators have shown, proactive interference dissipates with time (e.g., Loess and Waugh, 1967). Thus in the present experiments each pair of short-term memory tests can be treated independently. The procedure is described in more detail below. The first experiment was undertaken to determine whether this procedure would produce the general kind of forgetting function reported for short-term memory distractor experiments.

General Method

Materials. The items to be remembered in all experiments were consonant trigrams printed on slides. The trigrams used were not the same for all experiments. Within experiments, no letters were repeated within a trigram, and where letters were repeated across trigrams no letter appeared in consecutive trigrams or appeared more than once in the same trigram letter position.

Procedure. In each of five experiments, each subject in each condition received four pairs of short-term memory tests. A pair of short-term memory tests was defined as a trial. An interval of 2 to 3 minutes elapsed between trials to allow for the dissipation of proactive interference (Loess and Waugh, 1967). Although the events within a trial varied among the experiments and between conditions within the experiments, a trial always consisted of the following events defining the two, successive short-term memory tests. First, a ready signal; second, a consonant trigram read aloud once by the subject; third, a three-digit number, or a series of two-digit numbers presented throughout the retention interval which signaled arithmetic tasks to be performed aloud by the subject; fourth, a recall cue signaling the subject to recall the trigram orally. These events defined the first short-term memory test of the trial. Following an intertest interval these events were repeated for the second short-term memory test of that trial. The events of the two short-term memory tests were presented as a sequence of slides projected upon a transluscent screen by a Kodak carrousel projector and controlled by a tape timer. For all conditions in the five experiments, counterbalancing was such that each trigram appeared an equal number of times on each trial, and within a trial each trigram occurred equally in both the first and second short-term memory tests. Prior to an experiment, the subject received practice on the arithmetic task(s) he would encounter during the experiment.

EXPERIMENT I

Method

Subjects. The subjects were 48 University of Connecticut undergraduates who participated in the experiment as part of a course requirement. Each subject was allocated to one condition by order of appearance at the laboratory, with eight subjects per condition.

Materials. Beyond the constraints set in the general methodology the consonant-trigrams presented on short-term memory Tests 1 and 2 of a trial shared the same vowel sound, with the vowel sounds /i/ and /e/ alternating between trials, e.g., NLX and SFM as the pair of trigrams on one trial and PZB and VGC as the trigram pair on the next trial. The use of two different vowel sounds in this fashion was to protect further against the development of proactive interference across trials.

Procedure. Beyond the specifications in the general methodology the following trial events occurred: the ready signals were of 2 second duration; the recall intervals were of 10 second duration; the trigram presentations were of 2.5 second duration. On the first short-term memory test of a trial the retention interval was 5 seconds and the interval between the end of short-term memory Test 1 and the start of short-term memory Test 2 was 30 seconds. The second short-term memory test retention interval, which was the between-subjects variable, was 1, 3, 6, 10, 15, or 25 seconds. A three-digit number was present throughout a short-term memory test retention interval, and during the retention intervals the subjects counted backwards aloud at a self-paced rate by threes beginning at the indicated number.

Results and Discussion

For analysis, recall on a short-term memory test was scored such that a score of 1 was given if all three consonants were recalled in their correct order; if not, the recall was scored as zero. The recall data for the second short-term memory test of each trial were analyzed by means of a 6 (second short-term memory test retention interval) by 4 (trials) analysis of variance with the trial factor as a within-subject variable.

The trial order effect and the retention interval by trials interaction were not significant: $F(3,126)<1$, and $F(15,126)<1$, respectively. In addition, Test 1 recall on the four trials was respectively: .66, .85, .67, and .77. It would appear that the 2 to 3 minute intertrial interval and the shift in vowel class eliminated any systematic carryover effects from the preceding trial so that each two-test trial can be analytically treated as independent of previous trials.

Figure 1 illustrates the retention curve for short-term memory Test 2; the retention loss was significant, $F(5,42) = 12.9$, $p<0.001$. The retention curve of Figure 1 shows that forgetting reached a nonzero asymptote within approximately 10 seconds. The curve resembles an idealized curve suggested by Peterson (1966a). The observation of no further retention loss after 6 to 10 seconds in a short-term memory task is in concert with a very large number of experiments (e.g., Dillon and Reid, 1969; Merikle, 1968; Posner and Konick, 1966; Scheirer and Voss, 1969; Turvey, Brick, and Osborn, 1970b). It also agrees with the results of Baddeley and Scott (1971) who examined short-term memory performance in a situation in which there was no experimentally induced proactive interference.

EXPERIMENT II

The dissipation of proactive effects as a function of the elapsed time between successive short-term memory tests has been examined in several experiments (Cermak, 1970; Hopkins, Edwards, and Cook, 1972; Kincaid and Wickens, 1970; Leeming, 1968; Loess and Waugh, 1967). These experiments, however, do not

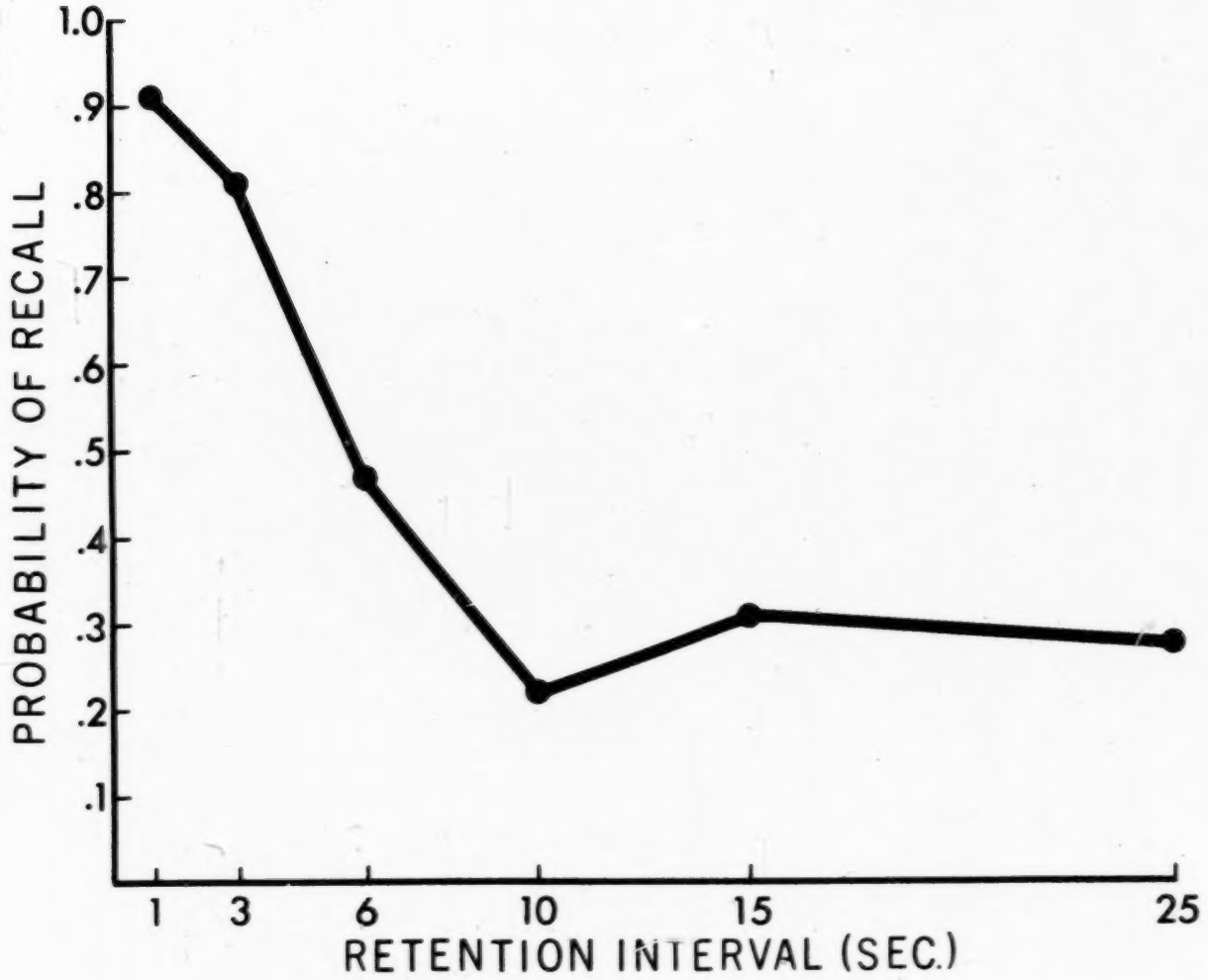


Figure 1: Test 2 retention in Experiment I as a function of retention interval.

directly concern the question of whether a reduction in proactive interference affects the short-term or long-term store component, or both components, of the short-term retention function. The second experiment, therefore, was conducted using the two-test procedure of Experiment I to examine in detail the locus of proactive interference effects on the short-term memory function.

Method

Subjects. The subjects were 120 University of Connecticut undergraduates who participated in the experiment as part of a course requirement. Each subject was allocated to one condition by order of appearance at the laboratory, with 12 subjects per condition.

Materials. Beyond the constraints set in the general methodology the consonant trigrams used in Experiment II were selected from the Scott and Baddeley (1971) norms. The trigrams were of moderate within-trigram acoustic similarity, ranging from 0.20 to 0.24, and of high association value, within the range of 0.71 to 0.83.

Procedure. The procedure was identical to that of Experiment I except that five short-term memory Test 2 retention intervals were employed (3, 4, 6, 8, and 30 seconds) and the intertest interval was either 0 or 40 seconds.

Results and Discussion

Recall was scored in the same fashion as in Experiment I. The short-term memory Test 1 recall probabilities on the four trials were .93, .88, .91, and .85 respectively, indicating an absence of any systematic carryover effects from preceding trials as demonstrated in Experiment I. The recall data for the second short-term memory tests of each trial were analyzed by means of a 5 (second short-term memory test retention interval) by 2 (intertest interval) by 4 (trials) analysis of variance with the trial factor as a within-subject variable. The main effect of retention interval was not significant, $F(4,110) = 1.141$, $p>0.05$, nor were any interactions, $P>0.05$, but the main effect of trials was significant, $F(3,330) = 6.48$, $P<0.001$. This latter effect was puzzling in that although it was highly significant it was far from systematic; across trials, the proportions of trigrams correctly recalled averaged across retention intervals were .71, .58, .71, and .83 respectively. The variables of major interest to the present experiment, the intertest intervals, proved to be significant, $F(1,110) = 5.05$, $P<0.05$, but comparisons across retention intervals showed that the two conditions differed only at the 4 second retention interval, $t(22) = 3.92$, $P<0.001$. Figure 2 illustrates the retention functions for the two intertest intervals. The significant difference between the two functions at the 4 second retention interval suggests the possibility that the short-term store component of short-term memory may be sensitive to proactive interference.

EXPERIMENT III

Experiment III was conducted to examine further the relation between proactive interference and the short-term memory function. Essentially, the third experiment was conducted as a replication of the second with three important variations: the use of a briefer retention interval to examine an earlier portion of the short-term memory function; the use of a subsidiary task during the intertest period; and the use of a longer intertest interval.

Figure 2

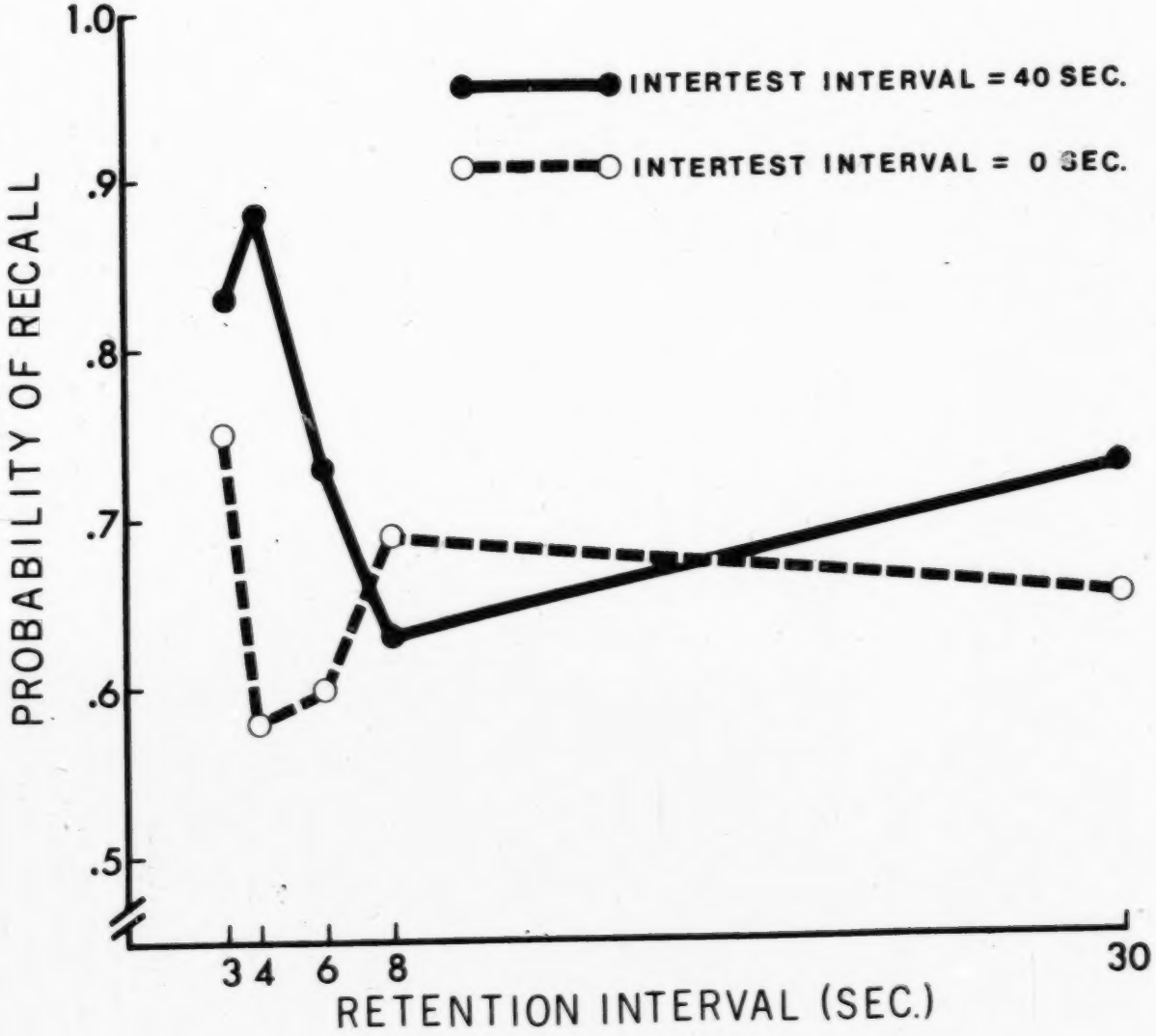


Figure 2: Test 2 retention in Experiment II as a function of retention interval and intertest interval.

Method

The subjects were 80 University of Connecticut undergraduates who participated in the experiment as part of a course requirement. Each subject was allocated to one condition by order of appearance at the laboratory, with eight subjects per condition.

The materials and procedure were identical in most respects to those of Experiment II. The few, but important, procedural differences were as follows: the trigram durations were 1.5 seconds; a 2 second short-term memory Test 2 retention interval was substituted for the 3 second interval of the previous experiment; and the intertest interval was either 0 or 70 seconds. In addition, subjects were required to engage in a series of addition problems during the 70 second intertest intervals.

Results and Discussion

For analysis, recall was scored in the usual manner. The Test 1 recall probabilities on the four trials were respectively: .81, .81, .78, and .79, indicating a lack of carryover effects from preceding trials. The recall data for the second short-term memory tests of each trial were analyzed by means of a 5 (second short-term memory test retention interval) by 2 (intertest interval) by 4 (trials) analysis of variance, with the trial factor as a within-subject variable. Neither the main effect of intertest interval nor the trials, not the interactions, were significant, $p>0.05$. On the other hand, the main effect of retention interval was significant, $F(4,70) = 3.98$, $P>0.01$. A t-test comparing performance in the 0 and 70 second intertest interval conditions at the 2 second retention interval was significant, $t(14) = 2.41$, $P<0.05$, as was a t-test on the 4 second retention interval data, $t(14) = 2.73$, $P<0.05$. The upper panel of Figure 3 illustrates the retention functions for the two intertest intervals.

As in Experiment II intertest interval appeared to have no effect upon the asymptotic portion of the retention curves; in both Experiments II and III the influence of the intertest interval was restricted to the early portion of retention. This result would appear to conflict with other results (Cermak, 1970; Kincaid and Wickens, 1970; Leeming, 1968; Loess and Waugh, 1967) which show intertest-interval effects at retention intervals ranging from 9 to 15 seconds. These latter findings might be interpreted as support for the view that proactive interference influences long-term store since it is the asymptotic portion of short-term memory functions which is affected (cf. Craik and Birtwistle, 1971). But quite to the contrary, the data presented in Figure 2 and the upper panel of Figure 3 show that the differences due to proactive interference occurred early in the retention interval. Estimates of short-term store or primary memory were obtained with the Kintsch et al. (1971) variation of the Waugh and Norman (1965) method using the average retention at 30 seconds as the estimate of long-term store. These estimates are plotted in the lower panel of Figure 3 and show quite clearly that the effect of intertest interval was upon the short-term store component of the retention function.

EXPERIMENT IV

The distractor technique for investigating retention over relatively short periods has been used (e.g., Posner and Konick, 1966; Posner and Rossman, 1965)

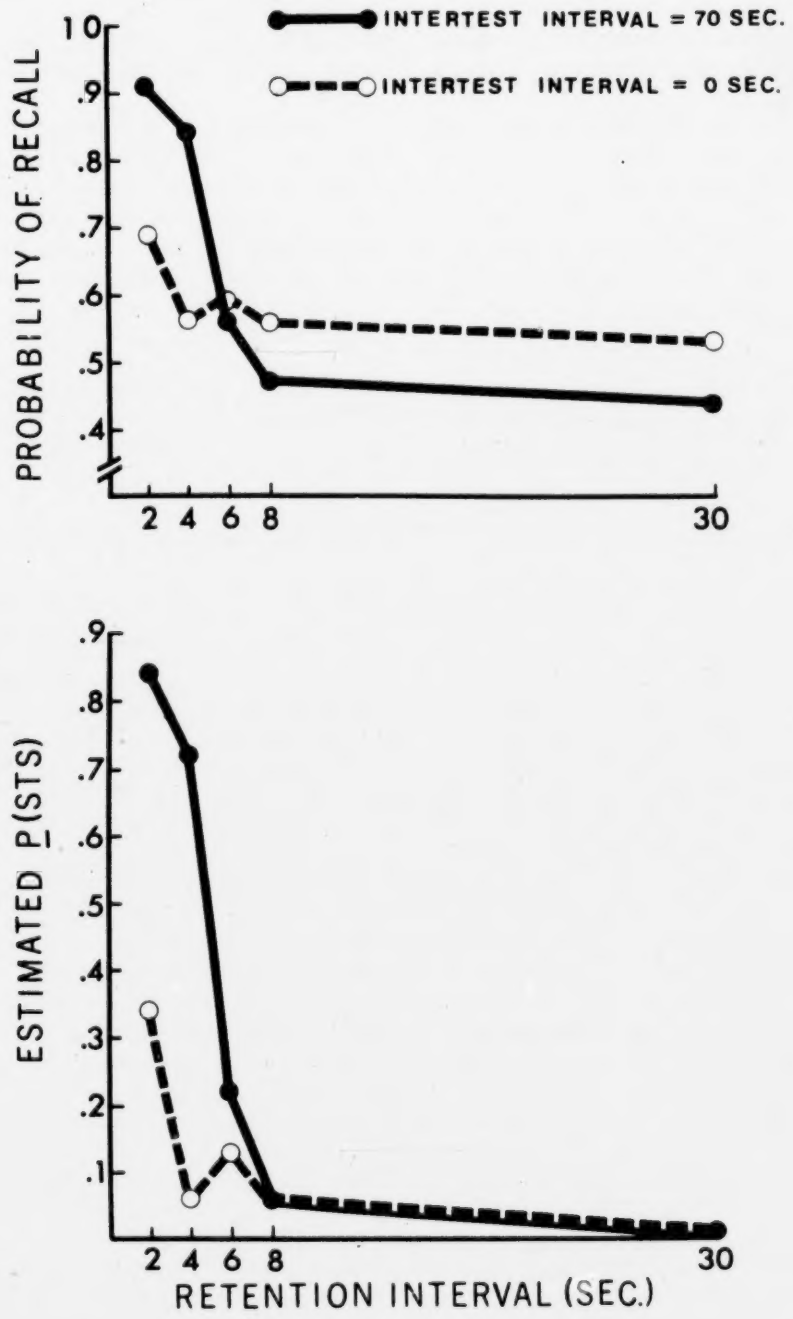


Figure 3: Test 2 retention (upper panel) and the estimated short-term store component (lower panel) on Test 2 in Experiment III as a function of retention interval and intertest interval.

to assess the effects of different levels of difficulty of a concurrently performed subsidiary task upon short-term retention. These experiments by Posner and his associates have demonstrated that the degree of forgetting is directly related to the amount of information reduction required by the subsidiary, or interpolated, task. More recently, Dillon and Reid (1969) have demonstrated that the major portion of forgetting occurs within 5 seconds of item's presentation and that the relation between interpolated task difficulty and the degree of forgetting is weighted heavily toward the earliest portion of the retention interval.

Experiment IV examined forgetting as a function of varying distractor complexity within the retention interval. The experiment of Dillon and Reid (1969) had used a continuous presentation of short-term memory tests, where each test was preceded by a short-term memory test whose interpolated task differed from test to test. By contrast, subjects in Experiment IV were presented a two short-term memory test series where conditions of the first test were consistent.

Method

Subjects. The subjects were 96 University of Connecticut undergraduates who participated in the experiment as part of a course requirement. Each subject was allocated to one condition by order of appearance at the laboratory, with eight subjects per condition.

Materials. The materials were identical to those of Experiment I.

Procedure. The procedure followed that of Experiment III except that: 1) The intertest interval was always 0 seconds. 2) The first short-term memory test retention interval was 10 seconds. 3) A series of two-digit numbers was presented during the second short-term memory test retention interval at a rate of one per 1.125 seconds. 4) There were four levels of second short-term memory test retention interval, 2.25, 4.5, 9, and 27 seconds, and four treatments during the second short-term memory test retention interval. The treatments consisted of: Condition DD, where a subject performed a difficult (D) task for the entire retention interval; Condition EE, where a subject performed an easy (E) task for the entire retention interval; Condition DE, where 4.5 seconds of the D task was followed by the E task for the remainder of the retention interval; Condition ED, where 4.5 seconds of the E task was followed by the D task for the remainder of the retention interval. With Conditions DE and ED only two values, 9 and 27 seconds, of the second short-term memory test retention interval were used. The D task was the summing aloud of the two digits of the two-digit numbers and the classifying aloud of that sum as odd or even; this task has an information reduction of 4.8 bits (cf. Dillon and Reid, 1969). The E task was the reading aloud of each two-digit number (zero bits information reduction), a task which, at the rate of one two-digit number per 1.125 seconds, has been shown to have a minimal effect upon retention (cf. Dillon and Reid, 1969). The E and D tasks in the DE and ED conditions were signaled to a subject by the underscoring of the two-digit numbers for the D task and no underscoring for the E task.

The sequences of two-digit numbers were constructed such that no pair of digits, or its reverse, occurred more than once within a trial; successive two-digit numbers with the same sum did not occur; the two digits in a pair were different; and a digit could occur in two successive pairs but not in the same position. Within these limitations the pairs of digits were chosen randomly. Before

the experiment subjects received practice on the arithmetic tasks that they would encounter during the experiment. One subject was dismissed for failure to achieve an acceptable level of performance on the interpolated tasks. The importance of good performance on the interpolated tasks during the experiment was emphasized to subjects.

Results and Discussion

For analysis, the dichotomous scoring procedure of the previous experiment was used. The recall data for the first of the two successive short-term memory tests were analyzed by means of a 4 (second short-term memory test retention interval) by 4 (second short-term memory test interpolated task) by 4 (trials) analysis of variance, with the trial factor as a within-subject variable. As expected, there were no significant main or interaction effects for the Test 1 data; however, recall performance was quite low with the mean proportion of items correctly recalled at 0.41.

A recall probability of .41 after a 10 second retention interval for conditions of zero, or at least minimal, proactive interference contrasts radically with the high retention reported in the comparable conditions of other experiments (Cofer and Davidson, 1968; Keppel and Underwood, 1962; Turvey, Brick, and Osborn, 1970a). On the other hand, significant short-term forgetting in the absence of proactive interference has been demonstrated by Baddeley and Scott (1971) (although not of the magnitude witnessed in the present experiment), and Petrusic and Dillon (1972) have drawn attention to the substantial Test 1 forgetting evident in a number of previous short-term memory experiments.

The short-term memory Test 2 recall data were analyzed in the same fashion as the Test 1 data. There were no within-subjects main or interaction effects. The retention function of each interpolated task condition is plotted in Figure 4. The main effect of short-term memory Test 2 retention interval was significant, $F(3,112) = 19.26$, $P<0.001$, as was the main effect of the interpolated task, $F(3,112) = 43.23$, $P<0.001$. A significant interaction between the second short-term memory test retention interval and the interpolated task was obtained, $F(9,112) = 3.16$, $P<0.005$. A Sheffe's multiple comparison revealed that at the 0.95 level of confidence there were no differences between the DD and DE conditions at any retention interval; recall at all retention intervals in the EE conditions was superior to recall in the DD and DE conditions; retention loss over time occurred only in the EE and ED conditions from 4.5 and 9 seconds. The ED condition at 9 and 27 seconds did not differ from the DD and DE conditions.

Forgetting was maximal within 9 seconds regardless of the interpolated task. In the condition where a D task followed the item to be remembered, forgetting was virtually complete within 2.25 seconds although continued forgetting may have been masked by a recall-probability floor effect. The relation among the recall probabilities of the conditions in the present experiment, with the one exception of the ED conditions, was analogous to the relation among the comparable conditions of the Dillon and Reid (1969) experiment. In that experiment, although the ED condition suffered more forgetting than the EE condition, the ED condition exhibited recall superior to the conditions where retention intervals were initiated by a D task. In the present experiment, however, the ED conditions were not significantly superior to the DD and DE conditions after the interpolated task shift. This failure to find a difference may have been due to a possible floor

Figure 4

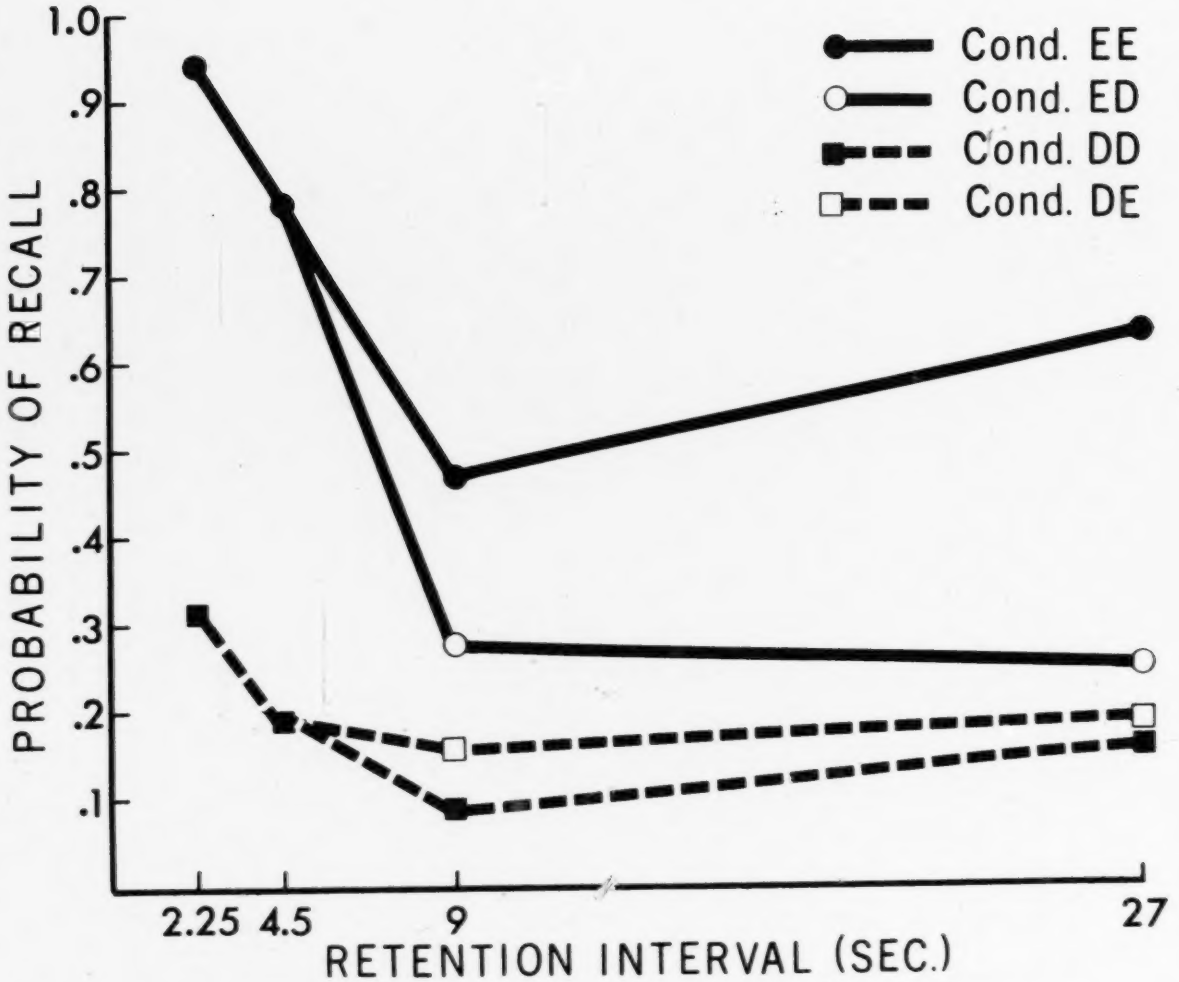


Figure 4: Test 2 retention in Experiment IV as a function of retention interval and Test 2 distractor task difficulty level.

effect in the DD and DE conditions. In any event, the finding of significant forgetting after the shift from E to D interpolated activity emphasized the underplaying by Dillon and Reid of the further retention loss accrued by the introduction of a D interpolated task. One possible interpretation of this finding is that the E task allows a subject to hold an item in memory (presumably short-term store) but consumes sufficient processing capacity to prohibit the item's registration in long-term store. When asked to recall early, the subject has the item available. However, if a D interpolated task now follows the E task, rehearsal as well as encoding processes are prohibited, and the item is lost. This issue is discussed more fully below.

A contingency analysis carried out to determine the relation between short-term memory Test 1 recall and short-term memory Test 2 recall proved to be significant, $x^2(1) = 10.48$, $P<0.005$: $P(\text{Cor 2}||\text{Cor 1}) = 0.27$; $P(\text{Cor 2}||\text{Incor 1}) = 0.44$. This means quite simply, that the probability of recalling an item was inversely related to the recall probability of the preceding item.

EXPERIMENT V

The argument that item-traces lose strength over time, and thus their potential for proactive interference, i.e., that proactive interference dissipates, has been noted above. The observed dissipation of proactive interference, however, may not be due to the weakening of the traces of preceding items but rather to improved cues for temporal discrimination (Baddeley and Scott, 1971), although Cermak (1970) for one has argued against the idea that temporal discrimination *per se* is responsible for the "dissipation" phenomenon. But the notion of continuously weakening traces of prior items, which purportedly underlies the dissipation of proactive interference, is not empirically supported. Maximum forgetting occurs within about 8 seconds after item's presentation and there is, on occasion, reminiscence (e.g., Peterson, 1966b; Scheirer and Voss, 1969; Turvey, Brick, and Osborn, 1970b). Therefore, why should further temporal separation result in a reduction of proactive interference if it is the case that recall probability reflects item strength, and that prior-item strength is related to proactive interference? Quite obviously the phenomenon of proactive interference dissipation cannot be understood solely in terms of trace attrition through time.

Since weakening of prior-item strength has been offered as an explanation of proactive interference dissipation, it becomes imperative to examine proactive interference effects where prior-item strength can be manipulated free from the potentially confounding influences of time. An alternative method of assessing the relation between proactive interference and short-term memory performance is one which uses interpolated task difficulty rather than time between tests as the variable regulating prior-item strength. Experiment V employs such a procedure, with the interval between the successive short-term memory tests of a trial held constant.

In addition, since task difficulty was manipulated in the retention interval of short-term memory Test 1, Experiment V afforded the opportunity to investigate the relation between interpolated task difficulty and forgetting in the absence of proactive interference.

Method

Subjects. The subjects were 192 University of Connecticut undergraduates who participated in the experiment as part of a course requirement. Each subject was allocated to one condition by order or appearance at the laboratory, with 16 subjects per condition.

Materials and procedure. The materials were identical to those of Experiment IV. The procedure was similar to that of Experiment IV with the following exceptions: 1) The second short-term memory test retention interval was 2, 8, or 24 seconds, and a three-digit number was present throughout the entire retention interval, during which a subject counted backwards aloud at a self-paced rate by threes beginning at the indicated number. 2) A series of two-digit numbers was presented during the first short-term memory test retention interval at the rate of one per 1.125 seconds, and the retention interval was 9 seconds. 3) During the first short-term memory test retention interval there were four treatments: Condition DD, where a subject performed the D task for the entire retention interval; Condition EE, where a subject performed the E task for the entire retention interval; Condition DE, where 4.5 seconds of the D task was followed by 4.5 seconds of the E task; Condition ED, where 4.5 seconds of the E task was followed by 4.5 seconds of the D task.

In an additional restriction beyond those set in Experiment IV, the sequences of two-digit numbers were constructed such that no pair of numbers, or its reverse, occurred more than once for a subject in the experiment. During practice on the arithmetic tasks five subjects were dismissed for failure to achieve an acceptable level of performance.

Results and Discussion

For analysis, recall was scored dichotomously as before. The recall data for the first short-term memory tests were analyzed by means of a 3 (second short-term memory test retention interval) by 4 (first short-term memory test interpolated task) by 4 (trials) analysis of variance, with the trial factor as a within-subject variable. Only the main effect of first short-term memory test interpolated task was significant, $F(3,148) = 25.94$, $P<0.001$. A Sheffe's multiple comparison revealed at the .95 level of confidence that DE = DD<ED<EE which yielded, respectively, recall proportions of 0.37, 0.46, 0.61, and 0.88. Thus, as in Experiment IV, retention losses occurred in the absence of proactive interference; with the exception of the EE condition, forgetting appeared to be greater than that found by Baddeley and Scott (1971). In addition, there was an inverse relation between interpolated task difficulty and probability of recall which paralleled the relation found in Experiment IV for second short-term memory tests.

The data for the second short-term memory tests were analyzed in the same manner as the first short-term memory tests. The main effect of Test 1 interpolated task was insignificant, $F(3,148) = 2.16$, $0.05<P<0.10$. The main effect of second short-term memory test retention interval was significant, $F(2,148) = 22.78$, $P<0.001$. The interaction between first short-term memory test interpolated task and second short-term memory test retention interval was significant, $F(6,148) = 2.39$, $P<0.05$. A Sheffe's multiple comparison revealed at the .95 level of confidence that at the 8 second retention interval $(DD + DE)/2 > (ED + EE)/2$.

These were, respectively, in proportion correctly recalled 0.48 and 0.31. Figure 5 illustrates the retention function of each treatment.

An effect in the expected direction of prior-item strength upon recall was suggested at the 8 second retention interval, but the effect disappeared by 24 seconds. Thus, as in Experiments II and III, it appears that proactive interference affects only the short-term store component of short-term forgetting. On the view that recall is dependent upon prior-item strength, then a clear and direct relation should exist in the present data between the difficulty level of the first short-term memory test interpolated task and the probability of short-term memory Test 2 recall, especially since it was demonstrated that interpolated task difficulty was directly related to recall on short-term memory Test 1. A contingency analysis carried out on the relation between first short-term memory test recall and second short-term memory test recall was significant, $x^2(1) = 7.22$, $P<0.01$; $P(\text{Cor 2}||\text{Cor 1}) = 0.39$; $P(\text{Cor 2}||\text{Incor 1}) = 0.50$. This direct relation between the degree of short-term memory Test 2 forgetting and short-term memory Test 1 recall was consistent with that found in Experiment IV.

GENERAL DISCUSSION

The fundamental feature of verbal memory under the conditions of the distractor paradigm is that forgetting reaches a maximum within a very brief period, usually of the order of 5 to 10 seconds. This characteristic of short-term retention is well exemplified in Figures 1 and 3 of this paper. While it has been argued often that short-term forgetting is due to the deleterious influence of previous material (e.g., Melton, 1963) current sentiment favors the more lenient view that traces of prior material contribute to this forgetting rather than being the necessary cause. Indeed, the present experiments (especially Experiment V) provide further evidence of memory impairment in the absence of proactive interference (Baddeley and Scott, 1971; Petrusic and Dillon, 1972). At all events, it is arguable that just preventing rehearsal will produce a significant degree of short-term forgetting.

The shape of the short-term forgetting function invites the interpretation that it represents two memorial systems, short-term store and long-term store, whose relative contributions to memory performance change with the lengthening of the retention period. The present experiments sought to determine whether the hypothesized short-term and long-term store components were affected independently or jointly by manipulation of proactive effects on the one hand and degree of rehearsal on the other. On the surface at least the results were singularly straightforward. Manipulating proactive interference, either by varying the recency or the rehearsal of prior material, affected only the short-term store component; in contrast, varying the complexity of the subsidiary task during the retention period with proactive effects held constant affected both short-term and long-term store. In short, short-term and long-term store were independently affected by proactive interference but jointly affected by demands on limited processing capacity.

There are currently two major views on the issue of how short-term store and long-term store are related: one is that the short-term store representation necessarily predates the long-term store representation, i.e., in due course material in short-term store is copied or transferred into long-term store; the other view is that short-term store is not a necessary precursor to long-term

Figure 5

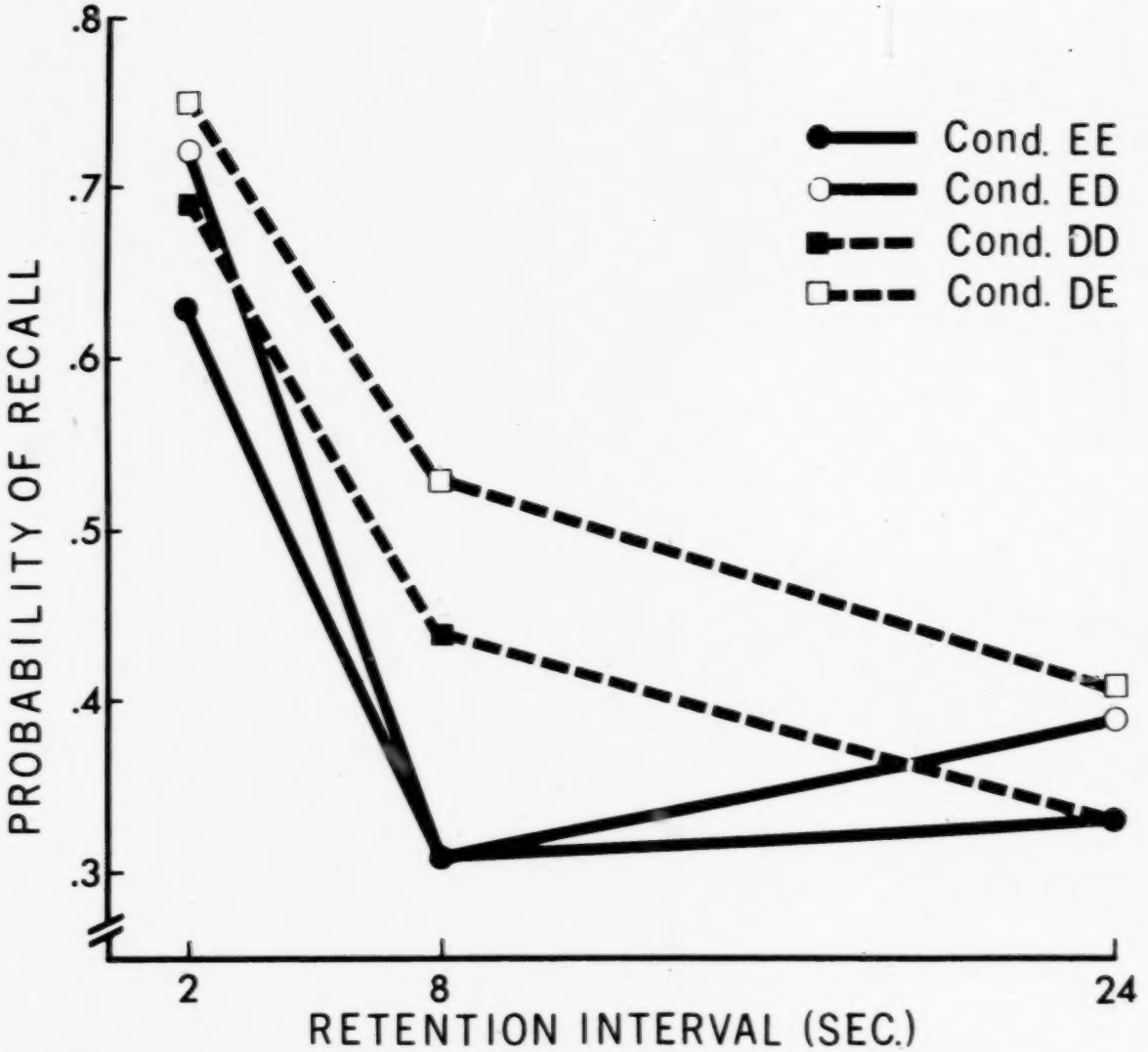


Figure 5: Test 2 retention in Experiment V as a function of retention interval and Test 1 distractor task difficulty level.

store and that memory representations are produced concurrently in both stores. Although the former has been the more commonly accepted of the two views (e.g., Atkinson and Shiffrin, 1968) there is reason to doubt its general validity. Smith, Barresi, and Gross (1971) and Peterson and Johnson (1971) have commented that according to this view we should expect that manipulations which facilitate or impair short-term store should lead to a proportionate facilitation or impairment in long-term store. Quite to the contrary, a number of experiments have demonstrated either a null or a negative correlation rather than a positive correlation between the two stores (e.g., Bartz and Salchi, 1970; Kintsch and Buschke, 1969; Kroll, Parkinson, and Parks, 1972; Peterson and Johnson, 1971; Smith, et al., 1971). In addition, Warrington and Shallice (1969, 1972) and Shallice and Warrington (1970) have described a patient whose longer-term retention of auditory material is under some circumstances relatively normal in spite of what appears to be a severe impediment in the retention of auditory material over brief periods. To these experiments we should add Experiments II, III, and V of the present series which similarly demonstrate that differences in recall in the initial period of a short-term memory function are not accompanied by proportionate differences at a later period.

In the face of these observations there are at least two ways to rescue the argument that the long-term store representation depends on short-term store. One is to assume that performance at early intervals is supported by one or more precategorical sensory buffers in addition to postcategorical short-term store. A precategorical acoustic buffer of the kind described by Crowder and Morton (1969) appears to play an important role in the initial retention of aurally presented material or of visually presented material which is vocalized by the subject (Kroll et al., 1972; Peterson and Johnson, 1972; Tell, 1971). This buffer, sometimes referred to as echoic storage, makes few, if any, demands on central processing capacity (Tell, 1971); is modality specific (cf. Crowder and Morton, 1969); and may have little or nothing to do with the fate of the longer-term representation (cf. Peterson and Johnson, 1972; Tell, 1971). Thus, a superiority manifest shortly after reception need not carry over to longer-term retention since its source may be the echo or some other kind of sensory buffer.

Alternatively, it could be argued that short-term store retrieval and long-term store retrieval are sensitive to different variables. On this view, a recall superiority in the early part of retention reflects a retrieval advantage rather than a storage advantage. Accordingly, superior recall at brief intervals need not mean, necessarily, superior registration in short-term store, and therefore, we should not always expect an early superiority to be manifest at later periods.

Whichever alternative is adopted to salvage the short-term store dependency view of long-term store it is quite evident that Experiment IV unequivocally supports the notion that long-term performance is directly related to short-term performance. The virtual elimination of the recency component in Conditions DE and DD of that experiment was paralleled by a practically nonexistent long-term store component. It would appear on this evidence that maintaining items in short-term store and registering items in long-term store <u>both</u> require a significant portion of a subject's limited central processing capacity (Broadbent, 1971; Moray, 1967; Posner, 1966). When that capacity is severely taxed, as it apparently is by the D task, then both the shorter- and the longer-term representations suffer. Assuming that this interpretation is correct, Experiment IV also suggests that in the immediate part of the retention interval priority in the allocation of

processing capacity or rehearsal is given to the sustaining of short-term store rather than to the elaboration of long-term store. Condition EE in Figure 4 shows that a significant short-term store component was present when the demands made upon rehearsal capacity were minimal. On the other hand, Condition ED suggests that very little long-term store strength accrued during the first 4.5 seconds of retention in Condition EE. Condition ED shows that when 4.5 seconds of the simpler task was followed by the D task for the remainder of the retention interval asymptotic retention was not significantly different from that of Conditions DD and DE. Although we cannot, of course, rule out the possibility of a floor effect obscuring real asymptote differences there remains sufficient reason to argue that during the initial period of Condition EE--at least the initial 4.5 seconds--the available processing capacity was allocated disproportionately between the two stores, with the short-term store as the major beneficiary.

By way of summary, Experiment IV makes the following comments relevant to the issue of how short-term store and long-term store are related. First, it argues that both the longer-term and the shorter-term representations supporting short-term memory performance are parasitic upon a limited processing capacity mechanism. Second, it suggests that the distribution of available capacity is initially biased in favor of sustaining the short-term representation. And third, it questions the view that the long-term and short-term representation are constructed simultaneously and exist independently.

With respect to proactive effects and short-term memory performance we are hard pressed to compare our results with those from other experiments which have investigated similar relations. In the first place, previous experiments which have looked at the effects of intertest interval have used only one retention interval and in each case it has been in excess of that in which we might expect to find a pronounced contribution from short-term store. Thus, in the experiments of Loess and Waugh (1967) the retention interval was 9 seconds; in Nield's (1969) experiment and the experiment of Kincaid and Wickens (1970) it was 10 seconds; Cermak (1970) examined retention after 12 seconds and the experiments of Hopkins et al. (1972) and Leeming (1969) both used a 15 second retention period. Since each of these experiments demonstrated that recall improved as the intertest interval was extended we should argue that these results support the view that long-term store is affected by proactive interference. On the other hand, the possibility remains that in at least some of these experiments short-term store was still quite prominent and that if longer retention periods had been examined the effects of intertest interval on short-term recall may have been insignificant.

Another difficulty facing the comparison of the present intertest interval results with those of previous experiments is that the previous experiments used multiple short-term memory tests. Presumably this was done either to build up maximal proactive effects prior to manipulating the intertest interval (e.g., Kincaid and Wickens, 1970) or to examine retention levels as a function of intertest interval under conditions of steady-state proactive interference (e.g., Loess and Waugh, 1967). In the present experiments only one test preceded the critical test. It could be argued, of course, that different determinants of forgetting arise when multiple short-term memory tests are used. In this light, it is encouraging to note that in Cermak's (1970) experiment there was no difference on Test 2 as a function of the preceding intertest interval (6 or 66 seconds), but a significant difference, owing to intertest interval, emerged later in short-term memory test series. Unfortunately, against this observation are the demonstrations by

Loess and Waugh (1967) and by Leeming (1968) of appreciable intertest-interval effects on the second of a series of short-term memory tests.

Whatever the reason for the difference between the present and previous findings it is quite clear, as noted above, that the present data contradict the view that proactive interference is limited to long-term store and that items in short-term store are impervious to the influence of prior traces (Craik and Birtwistle, 1971; Loess and Waugh, 1967). Given this contradiction, it should be argued perhaps that prior memory material affects the retention of subsequent material in several, and possibly quite different, ways. Broadbent (1971:342), for example, has commented that the principles governing the operation of proactive interference in the case of short-term forgetting are quite different from those we would expect from the investigation of longer-term forgetting. If we assume that within the short-term case alone there is more than one set of principles determining how prior material affects retention then we can understand why there might be discrepancies between our data and those of others, although, of course, we cannot account for them. In any event, the present data in conjunction with the data from other experiments suggest that the influences of prior memory material may be found in either the short-term or long-term component of short-term retention.

REFERENCES

Atkinson, R. C. and R. M. Shiffrin. (1968) Human memory: a proposed system and its control processes. In Advances in the Psychology of Learning and Motivation Research and Theory, Vol. 2, ed. by K. W. Spence and J. T. Spence. (New York: Academic Press) 89-195.

Baddeley, A. D. and D. Scott. (1971) Short-term forgetting in the absence of proactive interference. Quart. J. Psychol. 23, 275-283.

Bartz, W. H. and M. Salchi. (1970) Interference in short- and long-term memory. J. Exp. Psychol. 34, 380-382.

Broadbent, D. E. (1958) Perception and Communication. (New York: Pergamon).

Broadbent, D. E. (1971) Decision and Stress. (London: Academic Press).

Brown, J. (1958) Some tests of the decay theory of immediate memory. Quart. J. Exp. Psychol. 10, 12-21.

Cermak, L. S. (1970) Decay of interference as a function of the intertrial interval in short-term memory. J. Exp. Psychol. 84, 499-501.

Cofer, C. N. and E. H. Davidson. (1968) Proactive interference in STM for consonant units of two sizes. J. Verbal Learn. Verbal Behav. 7, 268-270.

Craik, F. I. M. and J. Birtwistle. (1971) Proactive inhibition in free recall. J. Exp. Psychol. 91, 120-123.

Crowder, R. G. and J. Morton. (1969) Precategorical acoustic storage (PAS). Percep. and Psychophys. 5, 365-373.

Dillon, R. F. and L. S. Reid. (1969) Short-term memory as a function of information processing during the retention interval. J. Exp. Psychol. 81, 261-269.

Hebb, D. O. (1949) The Organization of Behavior. (New York: Wiley).

Hopkins, R. H., R. E. Edwards, and C. L. Cook. (1972) The dissipation and release of proactive interference in a short-term memory task. Psychon. Sci. 27, 65-67.

Keppel, G. and B. J. Underwood. (1962) Proactive inhibition in short-term retention of single items. J. Verbal Learn. Verbal Behav. 1, 153-161.

Kincaid, J. P. and D. D. Wickens. (1970) Temporal gradient of release from proactive inhibition. J. Exp. Psychol. 86, 313-316.

Kintsch, W. and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.
Kintsch, W., E. J. Crothers, and C. C. Jorgensen. (1971) On the role of semantic processing in short-term retention. J. Exp. Psychol. 90, 96-101.
Kroll, N. E. A., S. R. Parkinson, and T. E. Parks. (1972) Sensory and active storage of compound visual and auditory stimuli. J. Exp. Psychol. 95, 32-38.
Leeming, F. C. (1968) Effects of association value of trigrams and ITI on short-term memory. Psychon. Sci. 11, 67-68.
Loess, H. and N. C. Waugh. (1967) Short-term memory and intertrial interval. J. Verbal Learn. Verbal Behav. 6, 455-460.
Melton, A. W. (1963) Implications of short-term memory for a general theory of memory. J. Verbal Learn. Verbal Behav. 2, 1-21.
Merikle, P. M. (1968) Unit size and interpolated task difficulty as determinants of short-term retention. J. Exp. Psychol. 77, 370-377.
Moray, N. (1967) Where is capacity limited? A survey and a model. Acta Psychol. (Amsterdam) 27, 84-92.
Nield, A. F. (1969) The effects of time and activity on the dissipation of proactive inhibition in short-term memory. Unpublished masters thesis, Ohio State University.
Peterson, L. R. (1966a) Short-term verbal memory and learning. Psychol. Rev. 73, 193-207.
Peterson, L. R. (1966b) Reminiscence in short-term retention. J. Exp. Psychol. 71, 115-118.
Peterson, L. R. and S. T. Johnson. (1971) Some effects of minimizing articulation on short-term retention. J. Verbal Learn. Verbal Behav. 10, 346-354.
Peterson, L. R. and M. J. Peterson. (1959) Short-term retention of individual items. J. Exp. Psychol. 58, 193-198.
Petrusic, W. M. and R. F. Dillon. (1972) Proactive interference in short-term recognition and recall memory. J. Exp. Psychol. 95, 412-418.
Posner, M. I. (1966) Components of skilled performance. Science 152, 1712-1718.
Posner, M. I. and A. F. Konick. (1966) On the role of interference in short-term retention. J. Exp. Psychol. 72, 221-231.
Posner, M. I. and E. Rossman. (1965) Effect of size and location of informational transforms upon short-term retention. J. Exp. Psychol. 70, 496-505.
Scheirer, C. J. and J. F. Voss. (1969) Reminiscence in short-term memory. J. Exp. Psychol. 80, 262-270.
Scott, D. and A. D. Baddeley. (1969) Acoustic confusability values for 1172 CCC trigrams. Psychon. Sci. 14, 189-190, 192.
Shallice, T. and E. K. Warrington. (1970) The independent functioning of verbal memory stores: a neuropsychological study. Quart. J. Exp. Psychol. 22, 261-273.
Smith, E. E., J. Barresi, and A. E. Gross. (1971) Imaginal versus verbal coding and the primary-secondary memory distinction. J. Verbal Learn. Verbal Behav. 10, 597-603.
Smith, T. L. (1896) On muscular memory. Amer. J. Psychol. 7, 453-490.
Smith, W. G. (1895) The relation of attention to memory. Mind 4, 47-73.
Tell, P. M. (1971) Influence of vocalization on short-term memory. J. Verbal Learn. Verbal Behav. 10, 149-156.
Turvey, M. T., P. Brick, and J. T. Osborn. (1970a) Proactive interference in short-term memory as a function of prior-item retention interval. Quart. J. Exp. Psychol. 22, 142-147.
Turvey, M. T., P. Brick, and J. T. Osborn. (1970b) Temporal course of proactive interference in short-term memory. Brit. J. Psychol. 61, 467-472.

Warrington, E. K. and T. Shallice. (1969) The selective impairment of auditory verbal short-term memory. Brain 92, 885-896.
Warrington, E. K. and T. Shallice. (1972) Neuropsychological evidence of visual storage in short-term memory tasks. Quart. J. Exp. Psychol. 24, 30-40.
Waugh, N. C. and D. A. Norman. (1965) Primary memory. Psychol. Rev. 72, 89-104.

On the Short-Term Retention of Serial, Tactile Stimuli*

Edie V. Sullivan[+] and M. T. Turvey[++]

In three experiments using the short-term memory distractor paradigm, Ss attempted to remember which three or four phalanges of the left hand had been stimulated and in which order. The experiments showed that forgetting increased as a function of trials, that such proactive effects could be eliminated by separating the successive trials by several minutes, that both verbal and nonverbal distractor tasks impaired retention, and that forgetting reached a maximum in approximately six seconds. All of these results concur with those generally obtained for the short-term retention of verbal material. In addition it was shown that the tactile recall was significantly poorer after an arithmetic distractor task presented visually than after the same task presented aurally. This result suggests an overlap between the mechanisms of tactile retention and the mechanisms of vision.

Most of what is known about the short-term memory (STM) performance of humans is based on experiments with verbal material, and it is often argued that STM is supported primarily by an acoustic-articulatory, or at least linguistic, code (e.g., Adams, 1967; Neisser, 1967; Sperling, 1963). There is evidence, however, that the short-term retention function for verbal material is affected by the input modality (e.g., Kroll, Parkinson, and Parks, 1972; Murdock, 1967). Moreover, it can be demonstrated that the short-term forgetting of verbal material is more severe if the rehearsal-preventing task uses the modality in which the memory material was perceived (e.g., Salzberg, Parks, Kroll, and Parkinson, 1971). Apparently an acoustic-articulatory code is not the only code [although it is unquestionably the more important code, (Conrad, 1972)] that supports the short-term retention of verbal items; nonverbal codes can also be involved.

It is of some importance to determine the characteristics of short-term nonverbal codes and their relation to the verbal codes which have played the central role in theories of memory dynamics (e.g., Atkinson and Shiffrin, 1968). As noted above, varying the modality in which verbal material is presented

---

*This report is based upon a thesis submitted for the MA degree by the first author under the supervision of the second author.

[+]University of Connecticut, Storrs.

[++]Haskins Laboratories, New Haven, Conn., and University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

affords one means for examining nonverbal STM codes. But such a procedure must suffer because where verbal material is involved the verbal STM code is likely to dominate and, therefore, likely to occlude some of the characteristics of the nonverbal representation. What is needed are STM tasks which use nonverbal stimuli and thereby prohibit or at least reduce the usage of verbal coding, and which, for purposes of comparison, are analogous to the STM tasks used to determine the properties of verbal retention.

Recent tactile analogues to verbal STM experiments (Gilson and Baddeley, 1969; Schurman, Bernstein, and Proctor, 1973; Sullivan and Turvey, 1972) have pointed to certain differences between the retention of verbal and tactile material over brief periods. First, although proactive effects are strongly evident in verbal STM tasks (Wickens, 1970), Sullivan and Turvey (1972) did not observe proactive effects in the retention of single tactile stimuli in the STM distractor paradigm (Brown, 1958; Peterson and Peterson, 1959). Second, in verbal STM experiments using the distractor paradigm, recall is perfect in the absence of a distractor, but in the tactile analogue, unfilled retention periods still yield forgetting. At the very least, the forgetting of tactile stimulation over unfilled retention periods suggests a memory-sustaining capability for tactile stimulation different from that available to verbal material.

The present series of experiments investigated several aspects of short-term tactile memory. The first experiment looked for the operation of proactive interference (PI) in a tactile memory task in which the spatial location and the temporal order of four tactile stimuli had to be retained. The first experiment also examined the effect of various retention-interval tasks on tactile retention. The second experiment questioned whether the forgetting function which characterizes verbal STM (cf. Peterson, 1966) similarly charterizes tactile STM. The third experiment examined the notion that at some level tactile retention involves the visual system (cf. Attneave and Benson, 1969)

GENERAL METHOD

The general approach was based, in part, on a series of experiments by Bliss and colleagues (e.g., Bliss, Crane, Mansfield, and Townsend, 1966). Stimulation was given to the underside of the 12 phalanges of the left hand (excluding the thumbs). Each was numbered. The index-finger phalanges, labeled in ink, were "1," "2," and "3" starting with the farthest from the palm of the hand to the most proximal. The remaining phalanges were labeled similarly, using the remaining numbers through to "12."

The E delivered either three or four stimuli per trial to approximately the center of the phalanges by means of a von Frey hair No. 8 inserted into the end of a 1-in cork. A von Frey hair is a nylon filament used as a pressure-controlling device (Carmon and Dyson, 1967). Following stimulation, E placed the von Frey hair into a 4-in plastic ring on the table in S's view so that when the recall cue was presented, the von Frey hair was readily available to S's right hand. Recall consisted of S attempting to stimulate, using the von Frey hair, the phalanges which E had stimulated at the outset of the trial. The Ss were instructed to move the pointer during the recall period until they felt confident about the accuracy of their reports (see Sanford, 1896:2; Sullivan and Turvey, 1972).

Each S sat to the right of E and of a table upon which S rested her left hand with the palm facing upwards so that the 12 numbered phalanges were in clear view of E. In all three experiments right-handed females served as Ss to minimize possible confounding resulting from handedness, laterality, or sex factors (Weinstein and Sersen, 1961); E, also a female, ran each S individually. During stimulus presentation, stimulus retention, and recall Ss were not allowed to look at their hands.

The Ss were exposed to at least one, but to not more than three, practice trials depending upon their individual needs to practice. When the Ss received four stimuli per trial, the recall scores were based on an eight-point scale, whereas a six-point scale provided the base when the Ss received three stimuli per trial. In either case, for each stimulus scored, one point was awarded for the recall of the correct spatial locale and one point for the recall of its proper temporal order. A Kodak carrousel slide projector controlled by a tape-timer provided procedural instructions and experimental timing for the Ss, with the exception of two conditions in Experiment III which used a Norelco cassette tape-recorder for instruction delivery.

For all three experiments, Ss were female, experimentally naive students from introductory psychology courses at the University of Connecticut who participated in the experiment in partial fulfillment of a course requirement.

EXPERIMENT I

While it is the case that verbal STM is sensitive to proactive effects the evidence available so far suggests, as noted above, that tactile STM is impervious to the influences of prior stimulation (e.g., Sullivan and Turvey, 1972). However, the absence of significant proactive effects in previous tactile STM experiments may have been due, in part, to the use of only a single point of stimulation for retention per trial. Indeed, detectable proactive effects would not be expected in verbal STM experiments where the memory material for each trial consisted of only one consonant (cf. Melton, 1963). Since PI is clearly evident in verbal STM when several items are presented for retention on each trial (e.g., Wickens, 1970) it was hypothesized that tactile STM would be similarly sensitive to proactive effects when the retention load per trial was several points, rather than one point, of stimulation. In addition, it was hypothesized that with several tactile stimuli to retain per trial, the activities performed by S during the retention period would exert a significant influence on recall accuracy. The evidence of prior research is equivocal concerning the effects of competing retention-interval activity on tactile STM. The retention-interval activity used in the previous experiments has been the arithmetic task generally employed in verbal STM experiments (cf. Peterson and Peterson, 1959). However tactile retention is fundamentally nonverbal. Therefore, we might suppose that nonverbal retention-interval activities, engaging processes similar to those involved in tactile retention, would prove to be more damaging than verbal activities (cf. Brooks, 1967, 1968). In any event, the present experiment looked at the effects of both verbal and nonverbal retention-interval activities on tactile STM.

Method

Subjects. Forty Ss were assigned, by order of appearance at the laboratory, to one of four conditions, with 10 Ss per condition.

Conditions. The four conditions were defined by four retention-interval tasks: rest, arithmetic task, tactile maze task, and visual maze task. For the rest condition Ss were informed that they should make an attempt during the retention interval to rehearse the points of stimulation "in their minds." In the arithmetic task condition Ss were shown a series of seven two-digit numbers, each exposed for 2 sec. Upon seeing a number, Ss were instructed to say it aloud, to add the two components of the number together, to tell E the answer, and to state whether the sum was odd or even. In the tactile maze condition Ss were given a random-choice, tactile maze. Without viewing the maze, Ss had to use the index finger of the right hand, i.e., the hand contralateral to that used for the memory task, to find their way from "start" to "finish." Upon hitting a blind alley they had to retrace their paths until they found a correct path. At each new trial, Ss were required to begin the maze at "start." They were told that E would inform them if they reached "finish." At the beginning of the experiment each S was allowed to glance at the maze merely to understand its general nature.

In the visual maze condition Ss were presented an 18-in by 12-in by 1/4-in peg board covering attached to a piece of styrofoam, 18-in by 12-in by 1-in. The visual maze was constructed so that its decision points were the same as those for the tactile maze. In this case, however, Ss were allowed to view the peg-board side of the maze while they inserted with the right hand a hooked peg into successive holes tracing right-angle paths to get from "start" to one of the holes at the top of the peg board. Diagonal paths were not allowed. In the styrofoam backing of the peg board were holes about 1/2-in deep designating the correct path. Therefore, if S selected a correct peg hole, the peg would enter the hole freely, while if S's choice were wrong, the peg would be stopped abruptly. As in the tactile maze condition, Ss were required to begin the maze at "start" at each new trial.

Procedure. Each S received five STM trials of the Brown-Peterson variety with the retention-interval activity held constant from trial to trial. On each trial, four successive tactile stimulations were given. A particular phalange could not be used more than twice for any given S on any given trial, and the same phalange could not be used on successive trials. The following phalanges, chosen randomly, were used twice for each S: 2, 4, 6, 8, 9, 10, 11, and 12. The quasi-randomized stimuli sets were counterbalanced across trials to avoid possible order effects. Five trials were considered sufficient to produce PI if it did exist; in verbal STM experiments, the significant recall decrement occurs over the first three trials (Loess, 1964; Turvey, Brick, and Osborn, 1970).

The Ss were asked to keep their eyes fixed on the slide projection area where their procedural instructions were shown. They were informed that they would receive a series of trials, all following the same procedure. The Ss were not told the number of trials. The events comprising a trial were as follows. The word "READY" typed in block letters in the center of the first slide appeared for 3 sec. A slide of red film paper followed for 4 sec, during which time E stimulated four of S's phalanges with the von Frey hair; S was not allowed to observe E's activities. For the next 14 sec, S engaged in one of the four experimental conditions. [In an earlier study, Sullivan and Turvey (1972) found that forgetting was maximal by a retention interval of 5 sec; therefore, an interval of 14 sec was used to insure forgetting within any given trial.] In the rest condition a series of seven, 2-sec blank slides were projected. In the arithmetic

task condition seven, 2-sec slides of two-digit numbers occurred. For the conditions using the mazes a 2-sec slide stating "START MAZE" was followed by six, 2-sec blank slides. These procedures stabilized conditions across Ss. The cue to recall was a 12-sec slide of the words "RECALL STIMULI." The end of the recall period was designated by a 5-sec blank slide which also served as the intertrial interval. An intertrial interval was used to insure the proper recording of S's recalls by E and E's preparation for the ensuring trial. According to verbal PI studies (Loess and Waugh, 1967; Nield, 1968) the intertrial interval between successive trials should be brief since PI effects vary inversely with the intertrial interval. Within the 12-sec recall period, Ss were asked to try to recall the points of stimulation in the order in which they were delivered and to do this with their right hand by means of the pointer (i.e., the von Frey hair) provided by E.

Results and Discussion

Recall on each trial was scored in the manner described above. An analysis of variance revealed that the effect of trials was significant, $F(4, 144) = 4.22$, $p<.01$; performance on later trials tended to be poorer than performance on earlier trials, and this decline in performance across trials defines PI. The effect of the retention-interval task was also significant, $F(3, 160) = 15.92$, $p<.001$, indicating that recall accuracy was dependent on what S did during the retention period. The analysis further revealed that there was no significant interaction between the two main effects ($F<1$). The functions relating retention-interval activity to trials are presented in Figure 1.

Experiment I suggests, therefore, that tactile STM, like its verbal counterpart, can be said to be sensitive to PI. It was argued above that proactive effects might be observed in tactile STM under conditions where several tactile locations had to be retained per trial. The outcome of Experiment I would appear to confirm this argument. However, an alternative view of the proactive effects of the present experiment is that they arose not because tactile STM is sensitive to PI but because Ss coded and stored the tactile stimuli verbally. Fortunately, there are several reasons to believe that verbal coding did not play a significant role in the present experiment. First, Ss, on questioning, infrequently reported the use of a verbal strategy for remembering the tactile stimuli. Indeed, verbal coding in the present tactile task is both difficult and time consuming, particularly given the rapid rate of stimulation of one phalange per second. (Readers are invited to try the task for themselves.) Second, if Ss were retaining the tactile input by verbal rehearsal then concurrent performance of the maze tasks should have disturbed retention less than concurrent performance of the arithmetic task. The former would have afforded more opportunity for repeating verbal descriptions than the latter. If anything, the data suggest (see Figure 1) that the maze tasks impaired retention more than the arithmetic task.

While the present experiment suggests that PI occurs in tactile STM it also suggests that PI is not essential for tactile forgetting since considerable forgetting occurred on the first trial of the series for all four conditions. Admittedly, the first trial was preceded by some practice of the tactile memory task and this may have provided a source of interference. However, the time elapsing between practice and presentation of Trial 1 was of the order of several minutes which should have been sufficient time for PI to dissipate. Significant

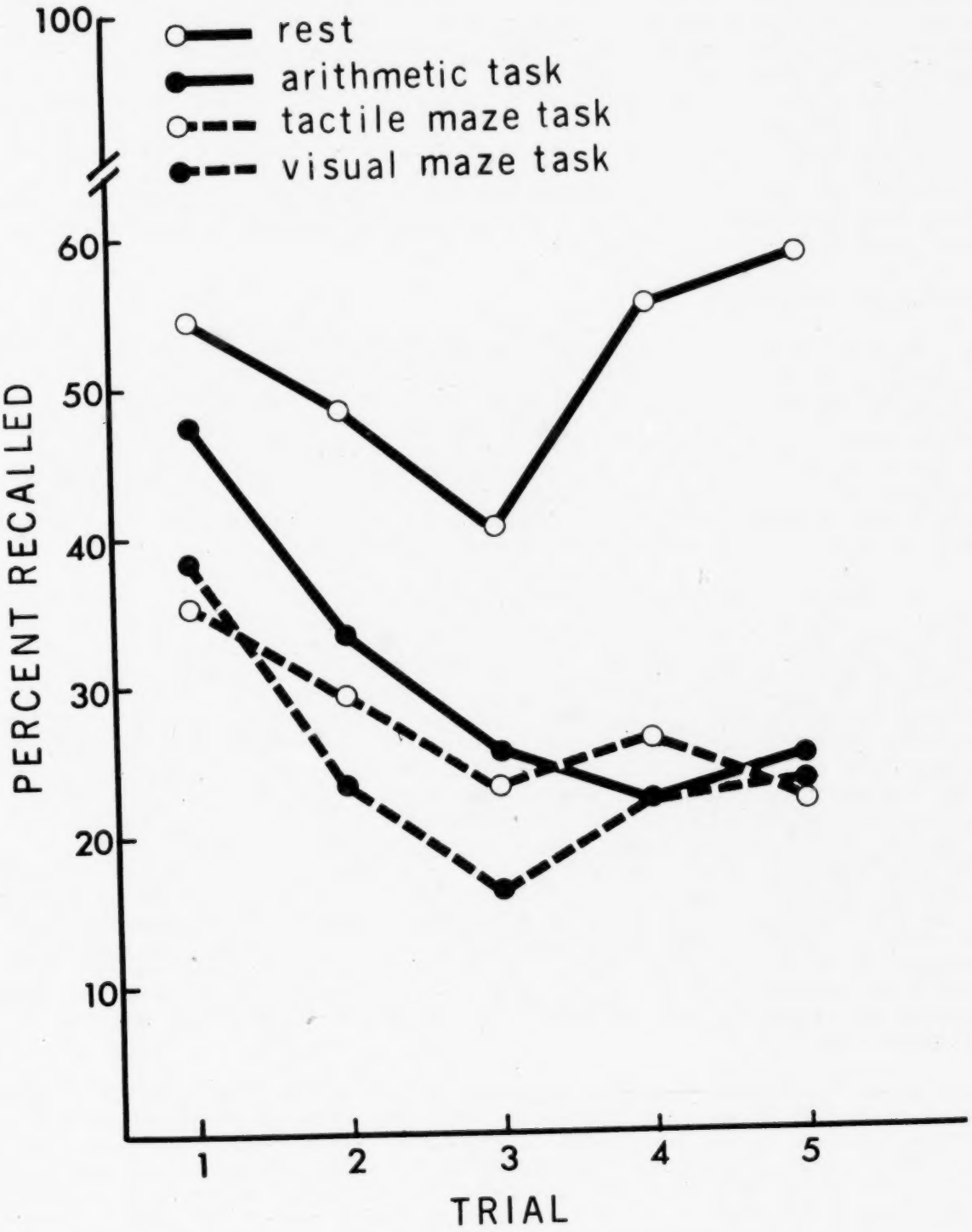


Figure 1: Short-term tactile recall in Experiment I as a function of trials with retention-interval activity as the curve parameter.

forgetting on the first trial of a series is similarly found in verbal STM experiments and, therefore, there is reason to argue that neither tactile nor verbal forgetting over brief periods need be precipitated by proactive effects (cf. Baddeley and Scott, 1971).

It is instructive to contrast the low level of recall in the rest condition of the present experiment with that which would be obtained if S was asked to retain four successive speech events, say consonants, over an unfilled interval. Performance on the latter task would be, of course, virtually perfect. The implication is that STM processes for ordered, discrete tactile locations are not especially efficient, at least not as efficient as those available to linguistic material. The source of this inefficiency could be in any one, two, or three of the stages of: initial coding, rehearsal, and recall.

The significant effect of the interpolated tasks in the present experiment suggests that rehearsal was important to the retention of four successive tactile stimuli. Of the three prior experiments on tactile STM, only one, that of Gilson and Baddeley (1969), found evidence for rehearsal in tactile STM, and that evidence was restricted, paradoxically, to the later retention periods. The positive finding of the present experiment was probably due to the use of several stimuli for retention per trial. The previous experiments (Gilson and Baddeley, 1969; Schurman et al., 1973; Sullivan and Turvey, 1972) had used a retention load of only one tactile location.

It was argued above that the tactile stimuli of the present experiment were probably not verbally coded. Therefore, "rehearsal" defined as inner speech would be inappropriate. Following Posner (1966), a more useful view of rehearsal is that it refers to a class of (unidentified) processes which sustain memory over brief periods, and which consume a portion of S's limited central processing capacity. The operations underlying the performance of the interpolated tasks would similarly require central processing capacity. Thus, in the present experiment the interpolated task, whether verbal or nonverbal, and the memory-sustaining operations on any one trial competed for S's limited attention; consequently, recall in the interpolated task conditions was notably poorer than in the rest condition. With respect to this discussion it is important to note that the smaller the memory load the less the demand on central processing capacity (Shulman and Greenberg, 1971).

EXPERIMENT II

The second experiment was designed to examine the short-term forgetting function for serial tactile stimuli. In addition, the experiment sought to determine whether the significant trials effect observed in Experiment I, could be eliminated if the interval between the trials was lengthened. In verbal STM experiments spacing the trials reduces PI (Kincaid and Wickens, 1970; Loess and Waugh, 1967; Nield, 1968).

Method

Subjects. Twenty-four Ss were allocated, on order of appearance at the laboratory, to one of three conditions, with eight Ss to each condition.

Procedure. The procedure followed that of Experiment I with the following exceptions. First, three phalanges, rather than four, were stimulated per trial

within a 3-sec period. Bliss et al. (1966) have shown that the immediate memory span in the present task is approximately 4.5 locations. By reducing the memory load to well within the memory span it was hoped that a higher level of performance than that of Experiment I would be obtained. Second, only one interpolated task, the arithmetic task, was used for all three conditions. This task was chosen because it was the most easily controlled of the three tasks used in Experiment I. Third, each S received four, rather than five, trials.

Four, three-stimuli sequences were used, with the constraint for each sequence that no phalange could be repeated, and horizontally or vertically adjacent phalanges could not be stimulated. The four sequences were balanced across Ss within a condition so that each sequence occurred twice in each trial position. A carrousel projector timed the events and the interval between the end of one trial, i.e., the end of the recall period, and the beginning of the next, was set at 2 min. Within the four-trial series each S received only one retention interval. The retention intervals used were 0, 6, and 28 sec, and these defined the three conditions of the experiment.

Results and Discussion

An analysis was initially conducted on the effect of trials. Averaged across Ss and across conditions, the recall accuracy on the four successive trials was: 53, 47, 48, and 47 percent. The analysis proved insignificant, suggesting that PI effects had been eliminated by the procedure of spacing the trials. The recall data were then averaged across trials within a condition and a simple analysis of variance revealed that the main effect of retention interval was significant, $F(2, 21) = 9.53$, $p<.005$. Recall in the 0, 6, and 28 sec conditions was, respectively, 67, 41, and 38 percent.

A typical interpretation of STM functions of the kind obtained in the present experiment is that they reflect two memory processes: a short-term component whose contribution to memory performance declines rapidly over the first few seconds and a relatively longer-term component whose contribution to performance is virtually invariant over brief retention periods (cf. Peterson, 1966). What is particularly important about the forgetting function obtained in the present experiment is that it compares favorably, both in its general form and in its point of maximal forgetting, with the forgetting functions obtained with verbal material in the distractor paradigm (Baddeley and Scott, 1971; Turvey and Weeks, 1973). The importance of this observation lies in its implication that common principles underlie the forgetting of verbal and nonverbal stimulation under conditions of filled retention periods. It should be noted that Sullivan and Turvey (1972) and Schurman et al. (1973) reported functions for tactile retention of single stimuli very similar to that reported here. In those experiments tactile forgetting reached an asymptote within 5 sec.

EXPERIMENT III

The point of departure for the third experiment was the notion that a visual-spatial code might underlie the retention of tactile location (Attneave and Benson, 1969). On this notion it was hypothesized that tactile STM would be more affected by an interpolated task presented visually than by the same task presented aurally. This hypothesis derives from experiments showing that visual activity is more antagonistic than aural activity to the construction and retention of spatial representations. Thus, Brooks (1967) showed that the spatial

coding of a message was poorer when the message was read than when it was heard. More recently, den Heyer and Barrett (1971) reported that the short-term retention of the location of letters in a visually presented matrix was selectively affected by a visual interpolated task, whereas the identity of the letters was selectively affected by an auditory task.

The third experiment was similar in its design to the two preceding experiments: Ss were required to retain a set of ordered tactile locations while performing an arithmetic task. The important difference was that the arithmetic task was presented either aurally, as before, or visually. The expected outcome was that tactile recall would be poorer in the latter case.

Method

Subjects. Each of 32 Ss was allocated to one of four conditions by order of appearance at the laboratory, with 8 Ss per condition. No S had participated in the previous experiments.

Procedure. The four conditions were: auditory arithmetic task (AAT), visual arithmetic task (VAT), auditory rest (AR), and visual rest (VR). During the retention interval of the AAT condition S heard a series of eight, two-digit numbers, each presented for 2 sec. In the VAT condition the same digit pairs were presented visually at the same rate. During the AR retention period, S heard a series of eight clicks at 2-sec intervals; during the VR retention period, S saw a series of eight blank slides, each of 2-sec duration. For both auditory conditions, the presentation of the three tactile stimuli per trial coincided with a steady-state tone; in the visual conditions, the tactile stimulation coincided with a red slide. No S was allowed to look at her left hand during the experiment. In addition, Ss in the auditory conditions kept their eyes closed during the retention interval in order to eliminate possible confounding visual effects.

Each of the four conditions used a 2-sec ready signal, a 3-sec tactile stimulation period, a 16-sec retention interval, a 12-sec recall period, and a 2-min intertrial interval. In each condition, each S was given eight trials, thus yielding 64 observations per condition. On completion of the experiment each S was questioned about the strategy she used to retain the tactile stimuli.

Results and Discussion

A representation of the recall performance in each condition is provided by Figure 2. An analysis of variance showed that there was a significant difference between performing the arithmetic task and resting, $F(1, 28) = 33.45$, $p<.001$. More important, the auditory/visual difference, and the interaction between modality and task, proved significant, $F(1, 28) = 5.07$, $p<.05$, and $F(1, 28) = 3.35$, $p<.025$, respectively. A Duncan's multiple range test revealed that at the $p = .05$ level, VR = AR > AAT > VAT.

The conclusion we would like to draw from this experiment is that tactile retention is impaired more by concurrent visual activity than by concurrent aural activity and, therefore, can be said to share processes more in common with vision than with audition. However, quite to the contrary, one could conclude that the difference between the VAT and AAT conditions was due to a

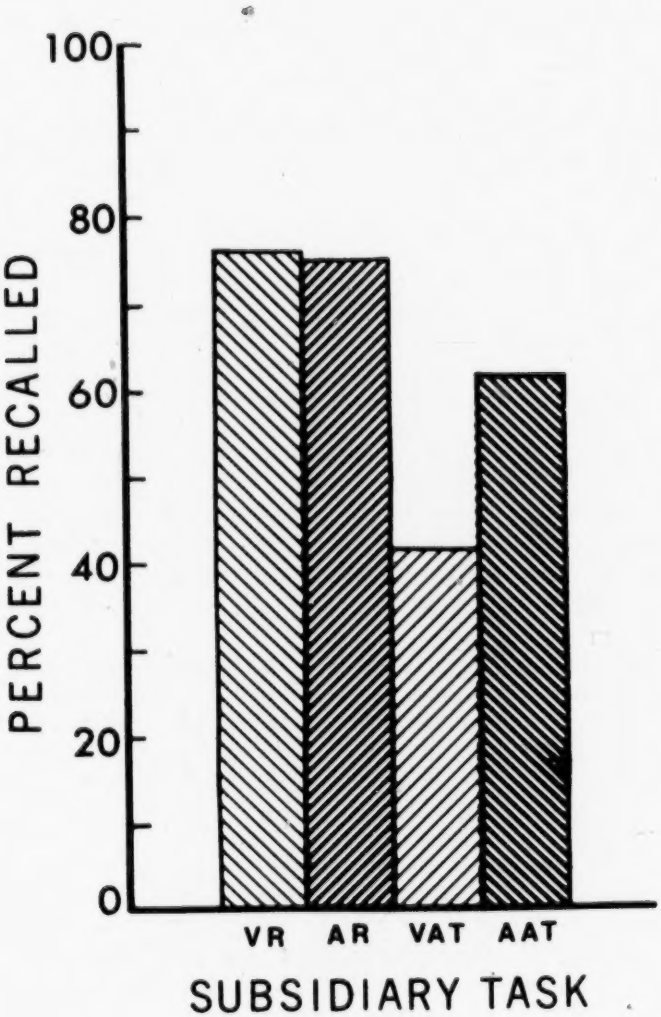


Figure 2: Short-term tactile recall in Experiment III as a function of the task performed during the retention period. VR = visual rest; AR = auditory rest; VAT = visual arithmetic task; AAT = auditory arithmetic task.

difference in the difficulty of the two interpolated tasks rather than to a difference in the modalities of their presentation. In short, the arithmetic task presented visually entailed operations over and above those of that task presented aurally, and these extra operations resulted in greater demands on S's limited processing capacity. This argument, however, is partially offset by the fact that performance on the arithmetic task was virtually identical for the two conditions. Performance on the digit-pair, addition and classification task, was scored as follows: one point for correctly repeating the two digits, one point for correctly adding them, and one point for reporting correctly the odd/even classification. On this measure the overall accuracy of performance in the VAT and AAT conditions was 90 percent and 94 percent, respectively. Of course, it could still be argued that while performance was equivalent in the two conditions the effort needed to achieve that level of performance in the VAT condition was greater than that needed in the AAT condition.

At best, the present data may be taken as only suggestive of a relation between the brief retention of spatially arranged tactile stimulation and visual processes. But data from other sources point in a similar direction: Brooks (1968: Experiment VII) showed that visualizing the spatial characteristics of a previously memorized diagram, and signaling those characteristics by means of tactually monitored movements, are, apparently, conflicting operations; Attneave and Benson (1968) demonstrated that learning responses to tactile stimuli delivered in fixed locations is better with unrestricted than with restricted vision.

On completion of the present experiment Ss were asked to report their strategies used to retain the tactile arrays. On the basis of their reports the Ss were divided into two groups: those who reported never using any form of a verbal code on any of the trials, and those who reported using some form of a verbal code on some of the trials. There were 16 Ss in the nonverbal group, and 15 Ss in the verbal group (E forgot to collect a report from one S). The average STM performance was equivalent for the two groups: nonverbal, 61 percent; verbal, 63 percent.

Verbal strategies consisted of trying to name either the finger stimulated or the part of the finger stimulated, i.e., top, middle, or bottom. In general, Ss reported that they found their verbal coding difficult, and for the most part inadequate. The nonverbal strategies involved either attempts to remember the "after-image" of the tactile stimulation, or attempts to image the spatial arrangement. There were also five Ss in the nonverbal group who were not aware of using any strategy whatsoever. In the VAT condition, there were three verbal coders and five nonverbal coders; in the AAT condition there were four verbal coders, four nonverbal coders. Thus, the significant difference between the two conditions was probably not due to differences in coding strategies.

By way of summary, the present series of experiments has shown that STM for serial, tactile stimuli is sensitive to the effects of PI; that forgetting of serial, tactile stimuli over brief periods is characterized by a rapid decline to asymptote within 6 sec; and that the retention of tactile stimulation might be dependent on mechanisms closely related to vision.

REFERENCES

Adams, J. A. (1967) Human Memory. (New York: McGraw-Hill).

Atkinson, R. C. and R. M. Shiffrin. (1968) Human memory: a proposed system and its control processes. In Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 2, ed. by K. W. Spence and J. T. Spence. (New York: Academic Press).

Attneave, F. and B. Benson. (1969) Spatial coding of tactual stimulation. J. Exp. Psychol. 81, 216-222.

Baddeley, A. D. and D. Scott. (1971) Word frequency and the unit sequence interference hypothesis in short-term memory. J. Verbal Learn. Verbal Behav. 10, 35-40.

Bliss, J. C., H. D. Crane, P. K. Mansfield, and J. T. Townsend. (1966) Information available in brief tactile presentation. Percep. Psychophys. 1, 273-283.

Brooks, L. R. (1967) The suppression of visualization by reading. Quart. J. Exp. Psychol. 19, 289-299.

Brooks, L. R. (1968) Spatial and verbal components of the act of recall. Canad. J. Psychol. 22, 349-368.

Brown, J. (1958) Some tests of the decay theory of immediate memory. Quart. J. Exp. Psychol. 10, 12-21.

Carmon, A. and J. A. Dyson. (1967) New instrumentation for research on tactile sensitivity and discrimination. Cortex 3, 406-418.

Conrad, R. (1972) Speech and reading. In Language by Ear and by Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Boston: MIT Press).

den Heyer, K. and B. Barrett. (1971) Selective loss of visual and verbal information in STM by means of visual and verbal interpolated tasks. Psychon. Sci. 25, 100-102

Gilson, E. Q. and A. D. Baddeley. (1969) Tactile short-term memory. Quart. J. Exp. Psychol. 21, 180-184.

Kincaid, J. P. and D. D. Wickens. (1970) Temporal gradient of release from proactive inhibition. J. Exp. Psychol. 86, 313-316.

Kroll, N. E. A., S. R. Parkinson, and T. E. Parks. (1972) Sensory and active storage of compound visual and auditory stimuli. J. Exp. Psychol. 95, 32-38.

Loess, H. (1964) Proactive inhibition in short-term memory. J. Verbal Learn. Verbal Behav. 3, 362-368.

Loess, H. and N. C. Waugh. (1967) Short-term memory and intertrial interval. J. Verbal Learn. Verbal Behav. 6, 455-460.

Melton, A. W. (1963) Implications of short-term memory for a general theory of memory. J. Verbal Learn. Verbal Behav. 2, 1-21.

Murdock, B. B., Jr. (1967) Auditory and visual stores in short-term memory. Acta Psychol. 27, 316-324.

Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).

Nield, A. F. (1968) The effects of time and activity on the dissipation of proactive inhibition in short-term memory. Unpublished Master's thesis, Ohio State University.

Peterson, L. R. (1966) Short-term verbal memory and learning. Psychol. Rev. 73, 193-207.

Peterson, L. R. and M. J. Peterson. (1959) Short-term retention of individual verbal items. J. Exp. Psychol. 58, 193-198.

Posner, M. I. (1966) Components of skilled performance. Science 152, 1712-1718.

Salzberg, P. N., T. E. Parks, N. E. A. Kroll, and S. R. Parkinson. (1971) Retroactive effects of phonemic similarity on short-term recall of visual and auditory stimuli. J. Exp. Psychol. 91, 43-46.

Sanford, E. C. (1896) A Course in Experimental Psychology. (Boston: D. C. Heath and Co.).
Schurman, D. L., I. H. Bernstein, and R. W. Proctor. (1973) Modality-specific short-term storage for pressure. Bull. Psychon. Soc. 1, 71-74.
Shulman, H. G. and S. N. Greenberg. (1971) Perceptual deficit due to division of attention between memory and perception. J. Exp. Psychol 88, 171-176.
Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.
Sullivan, E. V. and M. T. Turvey. (1972) Short-term retention of tactile stimulation. Quart. J. Exp. Psychol. 24, 253-261.
Turvey, M. T., P. Brick, and J. H. Osborn. (1970) Proactive interference in short-term memory as a function of prior-item retention interval. Quart. J. Exp. Psychol. 22, 142-147.
Turvey, M. T. and R. A. Weeks. (1973) Effects of proactive interference and rehearsal on the primary and secondary components of short-term retention. Haskins Laboratories Status Report on Speech Research SR-33 (this issue).
Weinstein, S. and E. A. Sersen. (1961) Tactual sensitivity as a function of handedness and laterality. J. Comp. Physiol. Psychol. 54, 665-669.
Wickens, D. D. (1970) Encoding categories of words: an empirical approach to meaning. Psychol. Rev. 77, 1-15.

Phonetic Activity in Reading: An Experiment with Kanji*

Donna Erickson,[+] Ignatius G. Mattingly,[+] and Michael T. Turvey[+]
Haskins Laboratories, New Haven

The Nature of Silent Reading

People have read aloud, and have been read to, for a very long time, but silent reading of substantial amounts of text has been until recently a fairly rare human accomplishment. In his Confessions (397 A.D.), Book VI, Augustine records his amazement at discovering his teacher, Ambrose, reading to himself: "When he was reading, his eye glided over the pages, and his heart searched out the sense, but his voice and tongue were at rest."

But silent reading is not necessarily silent. Indeed, for anyone who wants to understand this process, a basic problem is the nature, function, and extent of the concomitant phonetic activity. Such activity may conceivably be audible; or inaudible, but characterized by observable articulatory movements; or neuromotoric, without physical movement; or purely central. Some modern "silent" readers of English still move their lips and tongue or subvocalize, though such behavior is socially unacceptable in middle-class culture and is considered "immature" by teachers of reading. Many other readers who are deemed both mature and socially acceptable report that though they read silently, they are nevertheless aware as they read of a flow of phonetic imagery through the mind--a flow that has frequently been called "inner speech" (Huey, 1908; Goto, 1968). These readers, at least, will notice that a sentence such as "The rain in Spain falls mainly in the plain," has a phonetically bizarre pattern, although they have neither spoken the sentence nor heard it. With electromyographic techniques, attempts have been made with partial success to demonstrate a connection between inner speech and the neuromotor activity in the muscles that control the articulators (Sokolov, 1968).

Further evidence of phonetic activity during reading comes from experiments in visual information processing. One general finding is that when subjects are asked to perform a task requiring recall from short-term storage of letters or words read a few seconds before, the kinds of errors the subjects make suggest

---

*Some of the data and results in this paper have been reported earlier in Erickson, Mattingly, and Turvey (1972) and Kavanagh and Mattingly (1972:249).

[+]Also University of Connecticut, Storrs.

Acknowledgment: Special thanks are due to Hajime and Hatsue Hirose for their help in preparing the kanji stimuli.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1972)]

that this storage is phonetically organized (Conrad, 1972; Sperling, 1963). This kind of phonetic activity might well be purely central.

At any rate, some kind of phonetic activity often accompanies silent reading. But the function of this activity is not clear. Of course, the beginning reader has to acquire and exploit at least a passive knowledge of the spelling-to-sound rules of English in order to recognize in print a word that he has hitherto known only in its spoken form, and a mature reader uses these rules to assign a pronunciation to a printed word that he has not seen before. But even though the alphabet is in itself quasi-phonetic, English orthography, as Chomsky and Halle have insisted, is a better fit to the phonological than to the phonetic level of the language (Chomsky and Halle, 1968; Chomsky, 1970). That is, the spelling-to-sound correspondences exploit the linguistic rules that determine how the phonological representation of a morpheme is to be phonetically realized. Thus the regular plural morpheme /s/ is written s even though this morpheme is phonetically [əz], [z], or [s] according to phonological context, and /sign/ is sign in sign, resign, and signify, though pronounced differently in each of these three words. It might seem, therefore, that for a mature reader who is reading familiar material, phonetic activity is not only unnecessary but must pointlessly complicate the task. Since he is given in the printed text an essentially phonological representation and does not want to read it aloud, why can he not simply omit the reconstruction of the phonetic form of what he reads and deal only with syntactic and semantic reconstructions? Some people think that a genuinely skilled reader does in fact do just this, and there are some so-called "visual" readers who claim to go even further (Bloomfield, 1942; Bever and Bower, 1966; Bower, 1970). Like Ambrose, they glide over the pages, they report no inner speech, and they say that they "go directly to meaning" without any intermediate processing. It is difficult to know what to make of such claims, and we will return to them later. But at the very least, they seem to cast further doubt on the utility of phonetic activity during silent reading. Perhaps, as has often been suggested, phonetic activity in the mature reader is just an unfortunate vestige of early training to read aloud, and the less of it, the better.

But if phonetic activity is necessary to silent reading, we might still expect this to be true only in the case of alphabetic and syllabary orthographies, since these are the two classes of writing systems that overtly exploit the grammatical rules that generate phonetic representations. To be sure, the level of phonological abstractness varies substantially from one writing system to another. Finnish, Spanish, and Russian orthographies correspond in a fairly straightforward way with their respective phonetic systems, while the spelling-to-sound rules for English and French orthographies are quite complex. But regardless of the degree of abstractness, the rules of these writing systems parallel the phonological rules closely. Logographic writing, however, in which the symbols represent morphemes or morpheme compounds, might perhaps be processed quite differently. The reader of Chinese characters or Egyptian hieroglyphs might quite reasonably be expected to go directly to a morphemic representation, at least, if not to "meaning."

Yet a commonplace observation suggests that in at least one case, phonetic activity during the reading of symbols that have no overt phonetic structure at all, may be not only possible but perhaps, after all, necessary. Numerals have no such structure, yet almost all of us subvocalize vigorously when doing arithmetical computation on paper. Of course, subvocalization may be merely an

indication of the difficulty of the task, and hence no more relevant than subvocalization while tying a necktie or doing a jigsaw puzzle: we may be subvocalizing not the numbers themselves, but about the numbers.

But a result obtained by Klapp (1971) suggests more conclusively that the reading of numerals requires phonetic processing: the time taken to press a key to indicate that a pair of numbers were the same was measurably shorter for two-syllable numbers (e.g., 15 and 15) than for three-syllable numbers (e.g., 17 and 17). Still, it could be argued that the set of number words is too small to be interesting, or that, in Martin's (1972) phrase, numbers are "technological inventions thrust upon natural languages" and should not be the basis of general conclusions about the processing of linguistic symbols.

To sum up our difficulties, then, there is impressive evidence that some readers read silently with accompanying phonetic activity. Yet this activity is not obviously functional nor can we be sure that it is confined to writing systems that exploit phonological rules.

The Kanji Writing System

Let us turn now to consider more closely a writing system that, being logographic, would seem unlikely to necessitate phonetic activity: Japanese kanji.[1] The kanji characters were borrowed from the Chinese during several different epochs. They are logographic: a character represents a lexical morpheme without overt reference to its phonological or phonetic form. The morphemic value of a character is its "reading." Some characters have a Sino-Japanese reading derived from the Chinese reading of the character at the time of borrowing; some have a native Japanese reading; some have both kinds of reading; and some, having been borrowed more than once, have two or more readings of either or both kinds. Most of the characters consist of two elements: the "radical," which serves as a rough semantic classifier, and the "phonetic," so called (though misleadingly for our purposes) because it originally stood for a word similar in sound to the word represented by the character. Foreign words are spelled out with characters from an auxiliary writing system, the kata-kana, and a second kana system, the hiragana, is used in conjunction with the kanji for grammatical morphemes. In both kana systems there is a character for each vowel or consonant-vowel combination in Japanese; the kana scripts are quasi-phonetic. Moreover, the kana can represent adequately the vowel or consonant-vowel moras of which Japanese words are, phonologically speaking, composed.

Many Japanese with whom the authors are acquainted, and some Japanese linguists (e.g., Suzuki) claim to be visual readers of kanji. They say that they extract the meaning directly from the symbols without any phonological or phonetic activity. In this particular case, this claim is the more appealing for three reasons. First, even though the kanji were borrowed a long time ago, the original pictographic or ideographic reading of some of the most common characters still has some mnemonic value. For example, , originally , means "mountain."

---

[1] The brief account of Japanese writing given here is based on Palmer (1931), Suzuki (1962), and Martin (1972).

Second, spoken Japanese is inordinately homophonous, in part, as Martin (1972) says, because the Japanese ignored the contrasting tones of the Chinese morphemes they borrowed. There are extraordinarily many pairs of homophonous morphemes, and there are some homophonous sets with as many as 18 members. Moreover, there are numerous pairs whose members, unlike most homophonous pairs in English, are grammatically and semantically commutable, and must frequently give rise to real ambiguity. For example, the words for "national anthem" and "national flower" are homophones; so are the words for "integer" and "positive number." But sets of homophones have, in general, a different kanji character for each morpheme, and naive literate speakers seem to feel that the bond between a morpheme and its kanji is somehow more basic than the bond between the morpheme and its phonological form, the latter, according to Suzuki (1962) being "relegated to the background of our consciousness."

A third reason for attaching some weight to the claim that kanji are not read phonetically is provided by a study of Sasanuma and Fujimura (1971). They have reported that Japanese aphasics with apraxia of speech perform less well certain tasks requiring visual recognition and writing of kana than do aphasics without apraxia, while the two groups perform comparable tasks with kanji about equally well. This would seem to be consistent with the notion that some kind of phonetic activity is needed to read the kana but not to read the kanji.

An Experiment in the Recall of Kanji

In experiments with English-speaking subjects, it is generally found that confusions in short-term retention of alphabetic material are more often due to phonetic similarity between presented and recalled items than to visual or semantic similarity. It therefore is assumed that in retaining verbal material the hypothesized short-term storage system works primarily with a phonetic representation (Adams, 1967; Neisser, 1967). In the view of some students (e.g., Conrad, 1972; Liberman, Mattingly, and Turvey, 1972) this characteristic of short-term storage is of great importance to the reading process. However, in light of what we have just been saying about the unphonetic nature of the kanji and the attitude of native Japanese readers toward their language, we might expect to find that silent reading of kanji is not accompanied by phonetic activity of any kind, including in particular phonetic short-term memory. Such activity would seem to be not just superfluous, but a clear step backwards into homophony.

To test this hypothesis we made use of the probe short-term memory paradigm (Waugh and Norman, 1965). The procedure consists of presenting the subject a series of verbal items. On completion of the list one item from the list is presented as a "probe" and the subject's task is to identify the item that appeared immediately before the probe in the list: the "probed-for" item. The advantage of this procedure is that it allows for the separation of short-term storage effects from long-term storage effects. According to current memory theory, both the relatively transient representations in short-term storage and the relatively permanent representations in long-term storage support the retention of material over brief periods of time by short-term memory (Waugh and Norman, 1965; Atkinson and Shiffrin, 1968).

A recent experiment by Kintsch and Buschke (1969) is especially relevant to our present inquiry. These authors found that in a Waugh-Norman probe paradigm, a subject is less likely to retrieve the probed-for item from short-term storage if the items on the test list are homophones of one another. However, the

presence of homophones has no effect on retrieval from long-term storage. By contrast, if the items on the list are semantically similar to one another, retrieval from short-term storage is not affected, but retrieval from long-term storage is. We used this same tactic in an attempt to determine whether similar relations exist between phonetic and semantic similarity, and retrieval from short-term and long-term storage, when the material to be remembered is kanji characters.

Materials and design. Seven sets of 16 words were compiled. All the words used were concrete nouns of two moras and were written with one kanji containing 8-14 strokes. The nouns usually had only a single, Sino-Japanese reading but occasionally two readings were possible. Three of the seven sets were used for construction of practice lists to acquaint the subject with the procedure and to insure that in the actual experiment, practice effects would be relatively small. The other four sets were used to construct the experimental lists; they are given in Figures 1-4. One set (PS) consisted of eight pairs of words that were not only phonetically similar but in fact homophonous. In this set, all words ended in /ō/. A second set (SS) consisted of pairs of semantically similar words. A third set (OS) consisted of pairs of orthographically similar words: the radical elements in both members of a pair were the same, while the phonetic elements in both were different. A fourth set (NS) consisted of words that had no systematic similarity. Native informants were consulted informally about the compilation of all four sets. The SS set was particularly difficult to compile. The informants seemed to feel that the meaning of a word is so intimately related to its kanji that two words having different kanji can never be really synonymous, but only near in meaning to each other.

From each of these four 16-word sets 20 randomly permuted lists were constructed, subject to the following restrictions:

(a) One member of a pair of similar words could not immediately follow the other in a list, so that the probed and the probed-for words in a list were never similar.

(b) Across the 20 lists, each word served at least once, but not more than twice, as a probe and as a probed-for word.

(c) The probed-for word occurred twice each in list positions 3, 5, and 7, once each in positions 4 and 6, and four times each in positions 11, 13, and 15.

(d) In half the lists the probed-for word was 6 or 7 positions away from and in the other half of the lists it was 2 or 3 positions away from the word similar to it.

(e) In half the lists the probed-for word occurred earlier and in half occurred later the word similar to it.

Subjects. All subjects were members of a class in English for foreign students at the University of Hawaii. Ten were Japanese. Two other subjects, a Chinese and a Korean, were included because they could be expected to be familiar with the writing system but not with the Japanese language. [The Koreans have their own alphabet, but Chinese characters are traditionally used to write the numerous Chinese loanwords (Martin, 1972).]

PS List

	Reading	Kanji	Meaning
1	JŌ	城	castle
		嬢	girl
2	HŌ	砲	cannon
		報	report
3	BŌ	棒	stick
		帽	hat
4	CHŌ	腸	intestines
		蝶	butterfly
5	TŌ	燈	light
		陶	pottery
6	BYŌ	秒	second (time)
		病	illness
7	RYŌ	領	territory
		寮	dormitory
8	SŌ	僧	priest
		荘	inn

Figure I

SS List

	Meaning	Kanji	Reading
1	edge	縁	En, fuchi
		端	Tan, hashi
2	container	瓶	BIN
		壺	tsubo
3	room	室	SHITSU
		房	BŌ
4	grave	墓	haka
		塚	tsuka
5	district	郡	GUN
		県	KEN
6	evening	夜	yoru
		晩	BAN
7	bowl	椀	WAN
		鉢	hachi
8	mausoleum	陵	RYŌ
		廟	BYŌ

Figure 2

OS List

Radical		Kanji	Reading	Meaning
1	金	錠	JŌ	lock
		銃	JŪ	gun
2	女	姪	mei	niece
		婿	SEI	son-in-law
3	土	壕	GŌ	mound
		壇	DAN	platform
4	石	磯	iso	beach
		碑	HI	monument
5	月	脳	NŌ	brain
		膳	ZEN	tray
6	糸	紋	MON	crest
		線	SEN	line
7	木	棺	KAN	coffin
		校	KŌ	school
8	阝	院	IN	temple
		陣	JIN	battle line

Figure 3

NS List

	Kanji	Sino-Japanese Reading	Meaning
1	週	SHŪ	week
2	湾	WAN	bay
3	輪	RIN (wa)	wheel
4	像	ZŌ	statue
5	剣	KEN	sword
6	税	ZEI	tax
7	頬	KYŌ	cheek
8	綿	MEN	cotton
9	街	GAI	street (town)
10	談	DAN	story (talk)
11	隊	TAI	troop
12	塩	EN (shio)	salt
13	銀	GIN	silver
14	服	FUKU	cloth
15	斑	HAN	spot
16	穀	KOKU	kernel

Capital letters indicate Sino-Japanese readings; small letters, Japanese readings.

Figure 4

Procedure. Following a practice series, the 80 lists were presented on film strips to each subject in four blocks of 20, corresponding to the four similarity conditions. The words were presented at the rate of one per second, and the subjects read each word silently. One second after presentation of the last word on a list, the probe word (underlined) for that list was presented. The subjects then had to write down on a response sheet the kanji that had appeared immediately before the probe in the list. The position of the probed-for word was never the same on successive lists, nor did the same words or members of a similar pair appear as probed or probed-for words on successive lists. The four conditions were partially counterbalanced across the twelve subjects in a Latin-square design such that each condition appeared three times.

Results. Table 1 gives the recall probabilities averaged across the ten Japanese subjects for the kanji words of each condition as a function of serial position of the probed-for word. Inspection of Table 1 suggests that later items were recalled better than earlier items and that recall of PS words was poorer than recall of NS words, while recall of SS and OS words was not significantly poorer.

TABLE 1: Recall probability as a function of serial position for the Japanese subjects.

Condition	Average of Positions 3, 4, 5, 6, 7	Positions 11	13	15
NS	.13	.08	.22	.50
PS	.07	.00	.12	.35
SS	.10	.10	.18	.50
OS	.08	.10	.27	.52

We have already noted that performance in experiments where subjects are required to retain material over brief periods does not rely on short-term storage probability alone; rather, it is jointly determined by short-term and long-term storage probabilities (Waugh and Norman, 1965). Taking the view that the two storage systems are stochastically independent, Waugh and Norman proposed that in the probe short-term memory paradigm the probability of recalling an item is position i, $P(R_i)$ is:

$$P(R_i) = P(STS_i) + P(LTS) - P(STS_i)P(LTS)$$

where $P(STS_i)$ is the probability that item i is retained in short-term storage (STS) and $P(LTS)$ is the probability that it is retained in long-term storage (LTS). The assumption is made that $P(LTS)$ is independent of the position of an item in the list. $P(STS_i)$ is maximal for the most recent item, and decreases monotonically as a function of the distance from the end of the list, reaching zero after approximately 7-9 intervening items. If we regard the mean recall probability of items in positions 3, 4, 5, 6, and 7 as an estimate of $P(LTS)$ then

the equation above can be used to compute the STS component of the present data unconfounded by the LTS component.

Estimates of STS for positions 15, 13, and 11 were calculated in this fashion for each of the four conditions and are given in Figure 5 which also includes the corresponding LTS estimates. In Figure 5 we can see that retrieval from STS was significantly affected by phonetic similarity between the kanji but not by either semantic or orthographic similarity. This observation was borne out by a Wilcoxon matched-pairs test which showed a significant difference between PS and NS ($P<.05$) but no significant difference between SS and NS, or OS and NS. This is consistent with the finding of Kintsch and Buschke (1969) for short-term recall of English. A similar analysis conducted on the LTS estimates revealed no significant differences: perhaps because of the difficulties mentioned earlier in compiling semantically similar sets of kanji, we did not find for Japanese an effect of semantic similarity on long-term recall paralleling Kintsch and Buschke's for English.

The STS and LTS estimates averaged across the two non-Japanese subjects are given in Table 2. There are not enough subjects to provide results that

TABLE 2: Short-term and long-term storage estimates averaged for the Chinese and Korean subjects.

Condition	LTS Estimate	STS Estimates for Positions: 11	13	15
NS	.20	.03	.66	.84
PS	.15	.08	.39	1.00
SS	.15	.09	.28	.60
OS	.00	.30	.50	1.00

can compare in general with those from the Japanese, and in addition, a sequence effect may well be operating in the non-Japanese data. In any event, we do not know whether the relatively low SS score is significant, or why the LTS estimate for OS should be zero. Nor can we comment on the apparently higher level of STS performance achieved overall by the non-Japanese subjects. But the lack of any apparent PS effect for these two subjects offers some small assurance that the PS effect achieved with the Japanese subjects is not owing to some nonlinguistic artifact.

Discussion

In considering the significance of this experiment, we should keep in mind that its design was biased in several respects against the result that was obtained. The stimuli had no overt, systematic phonetic structure; the subjects, like most Japanese, did not believe that they used phonetic information in their ordinary reading of kanji text; linguistic considerations, as we have pointed out earlier, suggest that from a phonetic viewpoint Japanese is a less

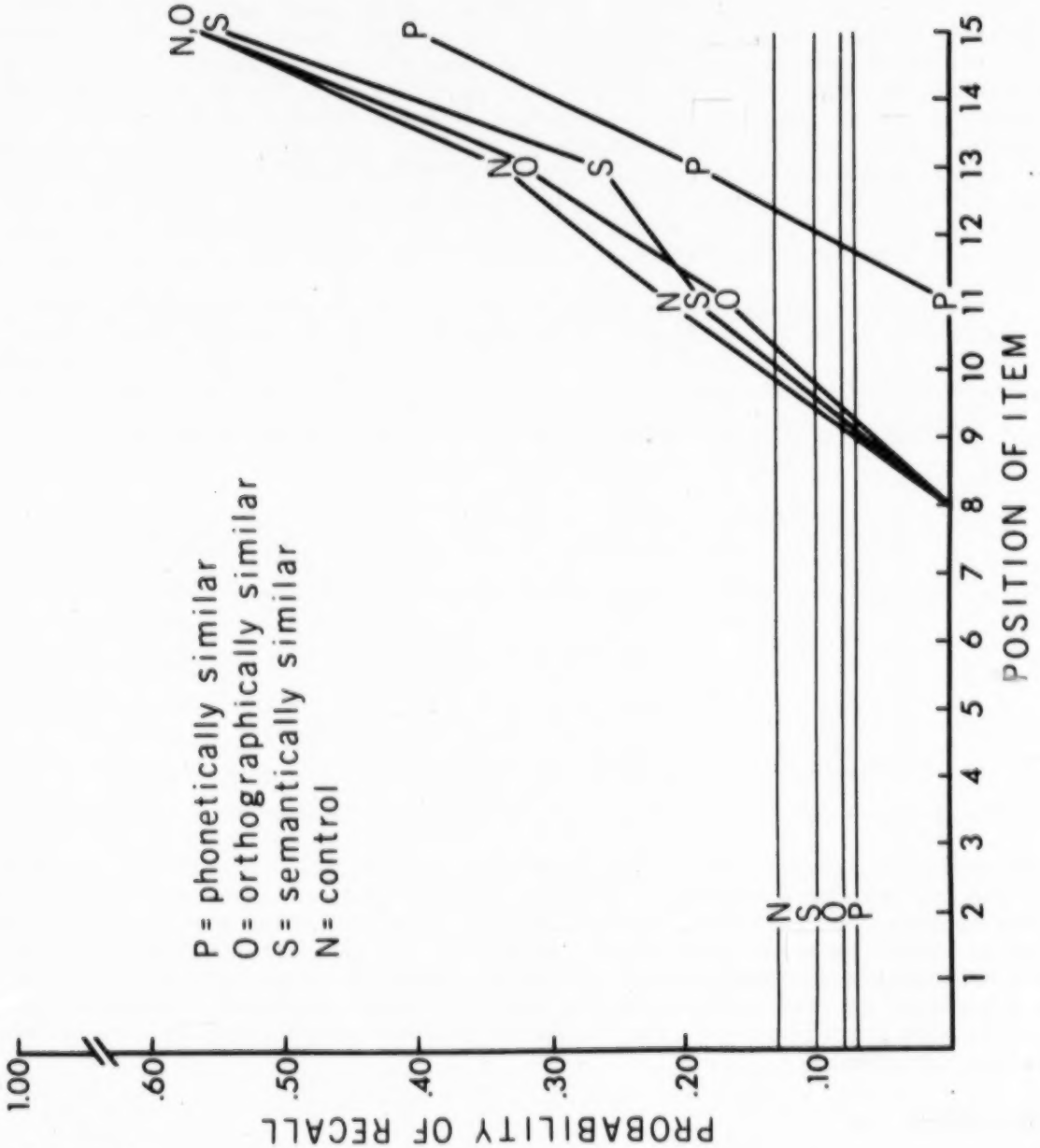

PROBABILITY OF RECALL
1.00
.60
.50
.40
.30
.20
.10
POSITION OF ITEM
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
P = phonetically similar
O = orthographically similar
S = semantically similar
N = control
N,O
S
P
N
S
O
P

Figure 5

than efficient means of communication; and the particular experimental task was such that a strategy that took account of the phonetic value of the characters would be likely to be a serious handicap with the PS set. Yet the significant difference between the PS and NS scores suggests that, despite these considerations, the subjects did resort to a phonetic strategy, and presumably they had a very good reason for doing so. We must try to establish what this reason could have been and what its implications are for an understanding of phonetic short-term storage, of the linguistic process, and of reading.

Phonetic short-term storage. The first observation that we can make concerns the conditions required to produce phonetic short-term storage effects in the visual processing of alphabetic material. There seem to be two possible views. The first is that the letters of the alphabet--quite aside from the lexical items they are used to represent--correspond with some degree of consistency to the phonetic values of a broad transcription, and since these correspondences are very familiar to the subjects, it is this primarily phonetic aspect of the letters that produces phonetic short-term storage effects, rather than any more abstract phonological, morphological, or syntactic factors. This view implies that phonetic activity is in principle readily separable from linguistic activity, and is all the more appealing because linguists, for their part, have traditionally insisted on the exclusion of phonetics from linguistics. It would follow that symbols that are not phonetic would evoke either some kind of nonphonetic short-term storage activity of their own, or no short-term storage activity at all. The second possibility is that, on the contrary, it is not the specifically phonetic but the more globally linguistic significance of alphabetic material that is important; phonetic short-term storage is only one component of a more general system for processing linguistic information, evoked by the task of reading the alphabetic stimuli. In this view, the nature of the short-term storage activity would depend on the system as well as the symbols, and phonetic activity evoked by nonphonetic symbols would not be surprising.

There is, of course, no way to distinguish these two possibilities clearly with alphabetic material. And it might be argued that the present experiment with kanji does not permit a perfectly clear distinction either, for the kanji do have phonetic values, in the sense that each symbol represents a lexical morpheme to which a phonetic value, i.e., its pronunciation, is assigned. Despite what has been said about the differences between logographic and alphabetic scripts, the reader of kanji controls a set of fixed, familiar correspondences between written symbols and spoken sounds just as the reader of the alphabet does. The difference is merely a matter of degree, the ensemble of units used for the short-term storage encoding being greater in the case of the kanji, and the units corresponding to longer phonetic events. Furthermore, it is irrelevant that the reader of kanji learns these symbols through the mediation of morphemes, since a similar observation might be made about many readers of the alphabet. Thus our result might be dismissed, though in a way that not only trivializes the experiment but also does violence to the intuitions of the native speaker of Japanese about his language and his writing system.

Fortunately, there is evidence of two sorts against the argument that there is no difference in kind between logographic and phonetic writing. First, there is Sasanuma and Fujimura's (1971) observation that the logographic kanji and the quasi-phonetic kana are differentially affected by brain damage. Whatever it may mean, this observation is not what we should expect if both writing systems were of essentially the same kind. Second, there is Wickelgren's (1966)

well-known demonstration of phonetic-feature confusions in the short-term retention of alphabetic text. Wickelgren's result suggests that the actual units of coding in short-term storage are to be clearly distinguished from the units symbolized by the characters of the writing system and that the short-term storage coding is a feature-by-feature representation and thus highly analytic, more so than any traditional writing system.

But if so, then there can be little question of coding the phonetic values of the kanji as monolithic units, and we can continue to regard kanji script (as linguists and native speakers have traditionally regarded it) as a morphologically based writing system. Since our experiment shows that the kanji nonetheless elicit phonetic short-term storage effects, we are compelled to consider seriously the second of the two explanations of such activity suggested above: the phonetic activity depends upon the essentially linguistic character of the experimental task. If this explanation is the correct one, it should be possible to demonstrate for any writing system in ordinary use the range of phonetic effects that have been found for alphabetic writing.

The experiments of Reicher (1969) and Wheeler (1970) support further and in a sense complement this view. These investigators found that under identical masking conditions a letter could be more accurately recognized if it was part of a four-letter word than if it was part of a four-letter anagram of this word. Apparently a subject knows that the sequence is meaningful before the process of identifying the letter shape is complete (cf. Turvey, in press). If the visual recognition of quasi-phonetic symbols invokes at the same time not only phonetic but also other linguistic processes, it is not surprising that the recall of symbols that are linguistic, but not overtly phonetic, should invoke phonetic as well as linguistic activity.

Linguistic process. Let us now consider what role phonetic short-term storage might play in the linguistic process. Following Chomsky (1965); Neisser (1967); and Liberman, Mattingly, and Turvey (1972), we regard what we shall call "primary linguistic activity" as a process of constructing a semantic representation and a phonetic representation. The phonetic representation is in a form appropriate for transmission; it is a specification for a certain pattern of articulatory activity being executed by a speaker and tracked by a listener. The semantic representation is in a form appropriate for storage; its relationship to long-term activity is parallel to (though as yet much less clearly understood than) the relationship of the phonetic representation to speech.

At present, we do not understand how the speaker-hearer constructs these representations. There are serious difficulties with the various simple models that have been proposed. We know a good deal about the form of the grammar that relates the two terminal representations of a given sentence, but we also know that we cannot map grammar into psychological process in any very straightforward way. We can tentatively infer from the grammar, however, that the construction of the two terminal representations involves several intermediate stages of recoding.

Phonetic short-term storage, as described by investigators of information processing, is clearly quite consistent with this general conception of primary linguistic activity. A basic fact that has emerged from information processing research is that the transformation of information is not instantaneous. It takes a significant (and with current techniques, measurable) amount of time,

for example, to register a visual "icon" and still more time to identify it as, say, a particular letter (Turvey, 1973). Therefore, we should expect that for primary linguistic activity in general, an appreciable period would also be required. But if so, a buffer or work-space of some kind is needed in which a representation of a sentence can be stored and updated during the course of linguistic processing. Phonetic short-term storage provides just what is called for. It is thus well motivated on psycholinguistic grounds, quite independently of its value in interpreting the results of information-processing experiments.

This view of the function of short-term storage bears on a question about primary linguistic activity that is very pertinent to reading. Since it has been suggested that there are several stages of recoding during this activity, can part of the process be short-circuited? Is it possible, for example, by presenting the processor with information in a form corresponding to some intermediate stage of recoding, to dispense with the phonetic buffer, and to use a buffer corresponding to this intermediate stage of recoding instead?

It is not at all unlikely that there should be a buffer storage associated with one or more of the intermediate stages of recoding. The prospect of a further proliferation of buffers should not dismay us; as George Miller has remarked (in Kavanagh and Mattingly, 1972:289), one would expect the nervous system to contain a good many buffers, just as the country contains a good many mailboxes. The question is open and obviously requires much further investigation. The outcome of the present experiment, however, gives no encouragement for this notion. The kanji correspond to a morphological or surface-structure representation, and so it might seem that part of the recoding has already been done for the reader of kanji. Yet phonetic activity is apparently necessary nonetheless. Our result is consistent with a very austere model of the linguistic processor in which there is no alternative buffer storage but only the phonetic buffer. Morphological information may require phonetic storage simply because no intermediate, morphological storage is available.

Yet we should be wary of concluding that there is absolutely no alternative to phonetic processing. As Conrad (1972) has shown, children who are congenitally and profoundly deaf develop nonphonetic strategies, no doubt very inefficient ones, for recall of written material, and they do learn to read, though in general poorly.

Reading. We turn now to the implications of our result for the nature of the reading process. We must, of course, proceed quite cautiously in extrapolating from a short-term memory experiment to any generalization about reading. The experiment, though simple in conception, required the subject to perform an apparently new and difficult task, practiced over just a few minutes. Reading ordinary connected text is a much more complex, yet far more familiar task, practiced over many years. But at least we can say that the outcome of the experiment is a further indication of a discrepancy already remarked upon. There is evidence of phonetic processing in the case of the Japanese reader just as there is in the case of an English reader. Those readers who say that they read without phonetic processing may well be reporting their subjective experience accurately, but they are probably processing phonetically, though quite unaware of doing so. Since we have just argued that this "unnecessary" processing is an essential part of primary linguistic activity, and since we might expect that reading, like performing in a verbal short-term memory experiment,

requires primary linguistic activity, this conclusion is not really surprising to us. But it seems to lead to a paradox. If phonetic short-term storage is so basic to primary linguistic activity, even when one is dealing with linguistic information in visual form, why are traditional writing systems not uniformly phonetic?

This is a knotty problem, and we can only suggest here the direction in which a solution should be sought. As Vygotsky (1934), Klima (1972), and Mattingly (1972) have suggested, it is possible to distinguish between primary linguistic activity itself and the speaker-hearer's awareness of this activity. If language were deliberately and consciously learned behavior, there would be little point to this distinction. But if we adopt the view of the generative grammarians that language is a very basic, highly elaborated cognitive process, acquired maturationally, our degree of linguistic awareness acquires a new significance. We should not necessarily expect to be any more aware of intermediate stages of linguistic process than we are of intermediate stages of other maturationally acquired behavior, such as walking or seeing. Yet we do seem to have a substantial, though far less than total linguistic awareness, and it is this awareness that, together with primary linguistic activity, forms the basis of various verbal skills that are forms of secondary linguistic activity. It is through the disciplined cultivation of linguistic awareness that speaker-hearers learn to speak a "secret" language like Pig Latin, or to compose verse that scans (Halle, 1970) or to transmit from generation to generation the traditional prose and poetry of a preliterate community. Such awareness, it has been suggested, is likewise the basis of reading and writing--literacy, too, is a secondary linguistic activity. We shall not attempt here to characterize the specific relationship between the secondary and the primary activities--indeed, if we could do so adequately, we would have arrived at an understanding of reading--but we may say that the skill of the reader consists in being able to control his primary linguistic activity in accordance with the written text, so that his awareness of this activity matches the symbols he sees.

Traditionally, of course, educators and even linguists have talked as if reading were in essence a form of speech perception in the visual modality. But in light of what is now known about the very peculiar nature of speech perception, this view seems to entail numerous contradictions, as Liberman (in Kavanagh, 1968) has pointed out.

If our view of reading as secondary linguistic activity is correct, we should expect that linguistic awareness, and not simply the requirements of primary linguistic activity, would determine the form of writing systems. Now the speaker-hearer's degree of linguistic awareness varies considerably over the range of primary linguistic activity; thus he is much more distinctly aware of the morphology of his language than he is of its phonology, and more aware of either of these than of its deep structure or its phonetics. This variation in linguistic awareness is clearly reflected in the history of writing systems (Gelb, 1963; Diringer, 1968). No culture has ever developed a writing system based on distinctive features or phrase-structure tree diagrams: such representations are useful only for linguistic description. At the other extreme, logographic scripts, which are morphologically based, are very ancient and have been invented independently many times. Sometimes those scripts have developed into syllabaries, which require a relatively modest phonological awareness. The alphabet was invented only once, not very long ago. As we have seen, alphabetic writing systems tend to be more nearly phonological than phonetic; rigorously

phonetic writing systems have indeed been proposed from time to time, but have never found general acceptance (Wilkins, 1668; Bell, 1867).

The alphabet, then, seems to be the least obvious of the traditional types of writing system. It became popular, however, because its practical advantages--the small size of the symbol inventory and the ease with which it could be adapted to new languages--in general more than compensated for its greater demand upon linguistic awareness. But this does not mean that alphabetic writing is the answer for every language. It is conceivable that the structure of a language, as it presents itself to the linguistic awareness of a native speaker, might be such as to make any quasi-phonetic writing system objectionable, despite the practical advantages. This would seem to be the case with Japanese. As we have already seen, the lexicon is filled with homophones. While this homophony probably does not interfere seriously with primary linguistic processing, it is quite conceivable that a writing system that failed to distinguish members of homophone sets would be highly distressing to the linguistic awareness of Japanese speaker-hearers. They are comfortable with a writing system that is morphologically rather than phonetically based, and understandably reluctant to adopt an alphabetic system that, because of the phonetic characteristics of the lexicon, would represent their language in the most ambiguous possible fashion. The well-meant attempts of reformers to replace the kanji with a Roman alphabet or with the syllabary kana (which do not resolve homophones either) have foundered on precisely this issue (Palmer, 1931).

A further implication for reading concerns the various kinds of phonetic activity discussed in our first section. The distinction we have drawn between primary linguistic activity, on one hand, and secondary linguistic activity dependent on linguistic awareness, on the other, leads us to conjecture that while phonetic short-term storage is a part of primary linguistic activity, the same is not necessarily true of inner speech or subvocalization, the presence or absence of which may vary with the reader, the degree of stress he is under, the writing system, and the content of the material being read. Rather, these latter phenomena seem to be associated with a certain kind of linguistic awareness.

If so, it may be possible to reconcile with our own result the finding of Sasanuma and Fujimura (1971), mentioned earlier, that aphasics with apraxia of speech, but not other aphasics, have special difficulties with kana, but not kanji. The interpretation of their finding suggested by these investigators runs counter to our own speculations: they argue that the kana spelling cannot be identified as a familiar representation of the word, but has to be interpreted as a phonetic code, which in turn can be related to the phonological representation of the lexical entry. A kanji transcription, on the other hand, can be directly identified with a word, bypassing the phonological interpretation, since each word (lexical entry) is independently associated with a kanji pattern as well as a phonological pattern. Thus the apraxic subjects are thought to have difficulty with the kana because they cannot bypass their damaged phonetics and phonology, as they can with the kanji. But there is a problem with this interpretation. If the group of aphasics that have relatively greater difficulty with kana suffer from a phonological impairment, we would surely expect them also to have relatively greater difficulty in speech perception. But the two groups differ only in speech production. Thus there is perhaps less reason to hypothesize a phonological impairment that is bypassed when reading kanji.

We suggest, very tentatively, an alternative explanation. Suppose that these subjects are apt to subvocalize (or engage in inner speech) when they read kana, but not when they read kanji. This would not only agree with the informal reports of normal Japanese readers, but would be consistent with the relationship between subvocalization and linguistic awareness that has been suggested above. The apraxic subjects could be expected to make the same kinds of performance errors in their subvocalizations as in their ordinary speech, and it would not be surprising if the resulting noisy feedback had an adverse effect on their performance of the experimental tasks involving kana. If this supposition is correct, we could then account for Sasanuma and Fujimura's result as reflecting a difference at the level of linguistic awareness, without hypothesizing that there are two fundamentally different kinds of primary linguistic activity associated with the two Japanese writing systems.

Let us now try to summarize our experimental result and the interpretation we have offered. We found that even though Japanese kanji characters have no overt phonetic structure, they are harder for native speakers to recall correctly when the set of characters to be remembered have phonetically similar readings. We took this to mean that phonetic short-term storage depends not on the type of writing system from which the stimuli are taken but on the linguistic nature of the task: phonetic short-term storage is necessary to primary linguistic activity. Such activity must require time, and phonetic short-term storage may serve as a kind of buffer in which sentences can be represented while linguistic processing goes on. The various types of writing system, on the other hand, reflect the fact that reading and writing are secondary processes that exploit the speaker-hearer's partial and varying awareness of his primary linguistic activity.

REFERENCES

Adams, J. A. (1967) Human Memory. (New York: McGraw-Hill).

Atkinson, R. C. and R. M. Shiffrin. (1968) Human memory: A proposed system and its control processes. In Psychology of Learning and Motivation: Advances in Research and Theory, Vol. 2, ed. by K. W. Spence and J. T. Spence. (New York: Academic Press).

Augustine. (397 A.D.) The Confessions, tr. by E. B. Pusey, Harvard Classics ed. (New York: P. F. Collier, 1933).

Bell, A. M. (1867) Visible Speech: The Science of Universal Alphabetics. (New York: Van Nostrand).

Bever, T. G. and T. G. Bower. (1966) How to read without listening. Project Literary Reports No. 6, 13-25.

Bloomfield, L. (1942) Linguistics and reading. Elementary English Review 19, 125-130; 183-186. [Also in A Leonard Bloomfield Anthology, ed. by C. F. Hockett. (Bloomington: Indiana University Press, 1970) 384-395.]

Bower, T. G. (1970) Reading by eye. In Basic Studies on Reading, ed. by H. Levin and J. Williams. (New York: Basic Books) 134-146.

Chomsky, N. (1965) Aspects of the Theory of Syntax. (Cambridge, Mass.: MIT Press).

Chomsky, N. (1970) Phonology and reading. In Basic Studies on Reading, ed. by H. Levin and J. Williams. (New York: Basic Books) 3-18.

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).

Conrad, R. (1972) Speech and reading. In Language By Ear and By Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 205-240.

Diringer, D. (1968) The Alphabet. (London: Hutchinson).
Erickson, D., I. G. Mattingly, and M. T. Turvey. (1972) Phonetic coding of kanji. J. Acoust. Soc. Amer. 52, 132.
Gelb, I. J. (1963) A Study of Writing. (Chicago: University of Chicago Press).
Goto, Hiromu. (1968) Studies on 'inner speech.' Part I: Reading and speech movement. Folia Psychiatrica et Neurologica Japanica 22, 65-77.
Halle, M. (1970) On metre and prosody. In Progress in Linguistics, ed. by M. Bierwisch and K. Heidolph. (The Hague: Mouton) 64-80.
Huey, E. B. (1908) The Psychology and Pedagogy of Reading. (New York: Macmillan). (Reprinted by MIT Press, Cambridge, Mass., 1968.)
Kavanagh, J. F., ed. (1968) Communicating by Language. The Reading Process. (Bethesda, Md.: National Institutes of Health).
Kavanagh, J. F. and I. G. Mattingly, eds. (1972) Language By Ear and By Eye. (Cambridge, Mass.: MIT Press).
Kintsch, W. and H. Buschke. (1969) Homophones and synonyms in short-term memory. J. Exp. Psychol. 80, 403-407.
Klapp, S. T. (1971) Implicit speech inferred from response latencies in same-different decisions. J. Exp. Psychol. 91, 262-267.
Klima, E. S. (1972) How alphabets might reflect language. In Language By Ear and By Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 57-80.
Liberman, A. M., I. G. Mattingly, and M. T. Turvey. (1972) Language codes and memory codes. In Coding Processes in Human Memory, ed. by A. W. Melton and E. Martin. (Washington: Winston) 307-334.
Martin, S. (1972) Nonalphabetic writing systems: Some observations. In Language By Ear and By Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 81-102.
Mattingly, I. G. (1972) Reading, the linguistic process, and linguistic awareness. In Language By Ear and By Eye, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press) 133-148.
Neisser, U. (1967) Cognitive Psychology. (New York: Appleton-Century-Crofts).
Palmer, H. E. (1931) Principles of Romanization. (Tokyo: Marozen).
Reicher, G. M. (1969) Perceptual recognition as a function of stimulus material. J. Exp. Psychol. 81, 275-280.
Sasanuma, S. and O. Fujimura. (1971) Selective impairment of phonetic and nonphonetic transcription of words in Japanese aphasic patients: Kana vs. kanji in visual recognition and writing. Cortex 7, 1-18.
Sokolov, A. N. (1968) Inner Speech and Thought. (Moscow, 1968). Tr. by G. T. Onischenko. (New York: Plenum, 1972).
Sperling, G. (1963) A model for visual memory tasks. Human Factors 5, 19-31.
Suzuki, Takao. (1962) A Semantic Analysis of Present Day Japanese with Particular Reference to the Role of Chinese Characters. (Tokyo: Keio Institute of Cultural and Linguistic Studies).
Turvey, M. T. (1973) On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. Psychol. Rev. 80, 1-52.
Turvey, M. T. (in press) Constructive theory, perceptual systems, and tacit knowledge. In Cognition and the Symbolic Processes, ed. by D. Palermo and W. Weimar. (Washington: Winston).
Vygotsky, L. S. (1934) Thought and Language. (Moscow-Leningrad, 1934). Tr. by E. Hanfmann and G. Vakar. (Cambridge, Mass.: MIT Press, 1962).
Waugh, N. C. and D. H. Norman. (1965) Primary memory. Psychol. Rev. 72, 89-104.

Wheeler, D. D. (1970) Processes in word recognition. Cog. Psychol. 1, 59-85.
Wickelgren, W. A. (1966) Distinctive features and errors in short-term memory for English consonants. J. Acoust. Soc. Amer. 39, 388-398.
Wilkins, J. (1668) An Essay Towards a Real Character and a Philosophical Language. (London: for S. Gellibrand).

Segmentation of the Spoken Word and Reading Acquisition*

Isabelle Y. Liberman+

There are many points of departure for investigators who are interested in reading. My colleagues and I at the University of Connecticut and Haskins Laboratories have begun with the fact that there are children who readily acquire the capacity to speak and listen to language but do not learn to read it. We have been concerned to ask, therefore, what is required in reading a language that is not required in speaking or listening to it.

The first answer that comes to mind, of course, is that reading requires visual identification of optical shapes. Since our concern here is with reading an alphabetic script, we may well ask whether the rapid identification of letters poses a major obstacle for children learning to read. The answer is that for most children, perception of letter shapes does not appear to be a serious problem. There is considerable agreement among investigators that by the end of the first year of school, even those children who make little further progress in learning to read generally show no significant difficulty in the visual identification of letters as such (Doehring, 1968; Kolers, 1972; Liberman, Shankweiler, Orlando, Harris, and Berti, 1971; Shankweiler, 1964; Vernon, 1960).

---

*Paper presented at the Symposium on Language and Perceptual Development in the Acquisition of Reading held at the biennial meeting of the Society for Research in Child Development, Philadelphia, Pa., 31 March 1973.

+University of Connecticut, Storrs. Although not a member of the Haskins Laboratories staff, the author has long been familiar with work done here. This paper reflects her interest and the Laboratories' interest in the implications of some of that work for an understanding of the reading process.

Acknowledgment: Much of the work reported in this paper was done in conjuction with Donald Shankweiler of the University of Connecticut. We are both deeply indebted to Alvin M. Liberman for many helpful suggestions. In the syllable-phoneme experiment, thanks are due to Bonnie Carter for aid in data collection and to F. William Fischer for his assistance in both data collection and statistical analysis. We are grateful also to Carol Fowler for her participation in all phases of the recent research reported here on consonants and vowels. Finally, all of us are indebted to Donald Libby, principal of Andover Elementary School, Andover, Conn., and to the teachers and pupils in that school without whose generous cooperation the research could not have been done at all.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

Beyond identification of letters, learning to read requires mastery of a system which maps the letters to units of speech. There is no evidence, however, that children have special difficulty in grasping the principle that letters stand for sounds. Indeed, children can generally make appropriate sounds in response to single letters, but are unable to proceed when they encounter the same letters in the context of words (Vernon, 1960).

A third possible source of difficulty is that the relation in English between spelling and language is often complex and irregular. But even when the items to be read are carefully chosen to include only those words which map the sound in a simple, consistent way and are part of the child's active vocabulary, many children continue to have difficulties (Savin, 1972).

What then are the real difficulties faced by the child in the early stages of reading acquisition? In this paper, I will explore one possible source of difficulty that has been recently proposed by us (Liberman, 1971; Shankweiler and Liberman, 1972) and other investigators (Elkonin, 1973; Klima, 1972; Mattingly, 1972)--that reading requires of the child an awareness of the structure of his language, an awareness that must be more explicit than is ever demanded in the ordinary course of listening and responding to speech. Since an alphabet is a cipher on the phonemes of a language, we should think that learning to decipher an alphabetically written word (as opposed to memorizing its visual configuration as may been done in learning so-called "sight" words) would require an ability to be quite explicit about the phonemic structure of the spoken word. For example, if the child is to map the printed word, "bat," which obviously has three letters, onto the spoken word which he already has in his lexicon, he must know that the spoken word also has three segments.

We suspect that this knowledge about the structure of the spoken word is not readily available to the child. Indeed, it appears not to have been readily available to the race. We know that an alphabetic method of writing, which rests upon an explicit phonemic analysis of the language, has been invented only once and is a comparatively recent development in the history of writing systems (Gelb, 1963). Syllabaries and logographic systems of writing, on the other hand, preceded the alphabet by thousands of years and have been invented independently several times. Of more immediate relevance to us is the evidence that children with reading disabilities often have difficulties even with spoken language when they are required to perform tasks demanding some degree of explicit segmentation of phonemic structure. These children are often reported, for example, to be deficient in rhyming, in recognizing that two different monosyllables may share the same first (or last) phonemic segment (Monroe, 1932), and, according to recent research (Savin, 1972), also in speaking Pig Latin, which demands a conscious shift of the initial phonemic segment to the final position in the word.

A third line of evidence suggesting that knowledge of phoneme structure is not readily available is provided by the behavior of reading-disabled children as observed by teachers who have worked with them (Johnson and Myklebust, 1967). Such a child will often demonstrate, as I have suggested earlier, that he can readily recover the phonemic segments in the ordinary course of speaking and listening. That is, he can respond appropriately to spoken words and to the objects to which they refer. Moreover, he can approximate the letter-to-sound correspondences. If he is asked, for example, to give the sound of the letter "b" he will say /bʌ/. For the sound for the letter "a" he will say /ae/ (though

this may give him more trouble, as we will discuss later). For the sound of the letter "t" he will say /tʌ/. But then if he is shown the printed word "bat" and asked to read it, he may give any one of a variety of incorrect responses (which I will deal with in more detail below in a discussion of error analysis). If he is then pressed to try to "sound it out," or otherwise to use what he knows about the letter-to-sound correspondences, he is likely to produce /bʌ/ /ae/ /tʌ/. At that point, he may be urged by the teacher to "say it faster," "put the sounds together," or, in the phrase commonly used, to "blend it." But no matter how fast he produces those sounds or how desperately he tries to put them together, he produces a nonsense word "buhatuh" containing five phonemic segments and not the word "bat," which has only three. Somehow, he cannot relate the three letters of the printed word to the three phonemic segments of the spoken word. It is as if he were not aware of the fact that the monosyllabic spoken word has three segments.

But why should it be so difficult for the child to become explicitly aware of phonemic segmentation? If, as has often been supposed, the sounds of speech bore a simple one-to-one relation to the phonemic structure just as the letters do (at least in the orthographically regular case), it would indeed be hard to see why phonemic analysis should pose special problems. That is, if there were in the word "bat" three acoustic segments, one for each of the three phonemes, then the segmentation of the word that is represented in its spelling would presumably be readily apparent.

However, as extensive research in speech perception has shown (Fant, 1962; A. Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967; Stevens, 1972), the segmentation of the acoustic signal does not correspond directly or in any easily determined way to the segmentation at the phonemic level. Moreover, this lack of correspondence does not arise because the sounds of the phonemes are merely linked together, as are the letters of the alphabet in cursive writing or as may be implied by the reading teacher who urges the child to blend "buhaguh" into a word that he knows. Instead, the phonemic segments are encoded at the acoustic level into essentially unitary sounds of approximately syllabic dimensions. In the case of "bat," for example, the initial and final consonants are folded into the medial vowel, with the result that information about successive segments is transmitted more or less simultaneously on the same parts of the sound (A. Liberman, 1970). In exactly that sense, the syllable "bat," which has three phonemic segments, has but one acoustic segment.

This is not to say that the phonemic elements are not real, but only that the relation between them and the sound is that of a very complex code, not a simple, one-to-one substitution cipher (A. Liberman et al., 1967). To recover the phonemic segments, to sort them out from the complex code, requires a correspondingly complex decoding process. In the normal course of perceiving speech these processes go on tacitly and automatically. To understand speech the listener need not be any more aware of the phonemic structure than he is of the rules of syntax.

Since the acoustic unit into which the phonemic elements are encoded is of approximately syllabic dimensions, one might suppose that the number of syllables (though not necessarily the exact location of the syllable boundaries) would be more readily apprehended than the phonemes. Syllable segmentation may be easier than phoneme segmentation for another reason as well. There are peaks of acoustic

energy (hence loudness) that correspond at least roughly to the vocalic nucleus of the syllable (Fletcher, 1929). Thus the syllable is acoustically marked, while the phoneme is not.

If syllabic segmentation is indeed easier, we might then have an explanation for the assertion (Makita, 1968) that the Japanese kana is readily mastered. The kana, one of the two Japanese writing systems, is approximately a syllabary. That is, most of the graphic symbols in the kana represent a syllable rather than a phoneme. There are separate symbols for ba, be, bi, bu, ga, ge, gi, gu, etc. Given the open syllable (CV) structure of the Japanese spoken language, the child therefore rarely needs to go below the level of the syllable in order to master the writing system. One might expect, further, that an orthography which represents each word with a different character (as is the case in Chinese ideographs or in the closely related Japanese kanji) would also not cause, in the beginning reader at least, the particular difficulties that arise in mastering the more analytic alphabetic system. Indirect evidence of the special burden imposed on the beginning reader by an alphabetic script can be found in the relative ease with which reading-disabled children learn kanji-like representations of language while being unable to break the alphabetic cipher (Rozin, Poritsky, and Sotsky, 1971). It is worth noting, in addition, that since the time of the ancient Greeks, methods of reading instruction have sporadically reflected the assumption on the part of educators that the phonemic structure of the language is more easily taught through the initial use of syllabic units (Mathews, 1966).

Though these considerations are suggestive, there has been no direct empirical test of the assumption that young children do, in fact, find it difficult to make an explicit phonemic analysis of the spoken word and that this ability comes later and is more difficult than syllabic analysis. My colleagues and I have undertaken in a recent experiment to provide such a test. For that purpose, we asked how well children can identify the number of phonemic segments in spoken words and how this compares with their ability to deal similarly with syllables.

The subjects were four-, five-, and six-year-olds in preschool, kindergarten, and first grade classes, respectively. They included 46 preschoolers, 49 kindergarteners, and 40 first graders. The unequal numbers arose from our plan to include all available children in the particular school at each grade level. Alphabetized class registers were used at each grade level to divide the children into the two experimental groups, one assigned to phoneme segmentation and the other to syllable segmentation. The level of intelligence of all the subjects was roughly assessed by means of the Goodenough Draw-a-Person Test. Two-way analyses of variance performed on the Goodenough DAP scores revealed no significant differences in IQ, either across tasks or across grade levels. The mean chronological ages of the two task groups were also not significantly different. Therefore, any performance differences in the two types of segmentation can reasonably be taken to reflect differences in the difficulty of the two tasks.

The procedure was in the form of a tapping game. The child was required to repeat a word or sound spoken by the examiner and to indicate, by tapping a wooden dowel on the table, the number (from one to three) of the segments (phonemes in one group, syllables in the other) in the stimulus items. Four sets of training trials containing three items each were given to both groups. The test trials, which followed the four sets of training trials, consisted of 42 randomly assorted individual items of one, two, and three segments which were

presented without prior demonstration and corrected, as needed, immediately after the child's response. Testing was continued through all 42 items or until the child reached criterion of tapping six consecutive trials correctly without demonstration. Instructions given to the two experimental groups at all three age levels were identical except that the training and test items involved phonemic segmentation in one group and syllabic segmentation in the other. All the children were tested close to the end of the school year.

The results showed in many ways that the test items were more readily segmented into syllables than into phonemes. In the first place, the number of children who were able to reach criterion was markedly greater in the syllable group than in the phoneme group, whatever the grade level. At age four, none of the children could segment by phonemes, while nearly half could meet the stringent criterion with the syllables. Ability to perform phoneme segmentation successfully did not appear at all until age five, and then it was demonstrated by only 17% of the children. In contrast, almost half of the children at that age could segment syllabically. Even at age six, only 70% succeeded in phoneme segmentation, while 90% were successful in the syllable task.

The contrast in difficulty can also be seen in terms of the number of children who achieved criterion level in six trials, which, under the procedures of the experiment, was the minimum number possible. For the children who worked at the syllable tasks, the percentage reaching criterion in the minimum time increased steadily over the three age levels. It was 7% at age four, 16% at age five, and 50% at age six. In striking contrast to this, we find that in the phoneme group, no child at any grade level attained the criterion in the minimum time. An analysis of variance which assessed the contribution of task and grade found that these main effects were highly significant, with a p level of less than .001.

We cannot judge from this experiment to what degree the measured increases in phoneme segmentation with age represent maturational changes and to what extent they may reflect the effects of instruction in reading. We would guess that the sharp increase from 17% at age five to 70% at age six in the number of children passing the phoneme task is probably due in large part to the intensive concentration on reading and readiness activities in the first grade. The possibility that these changes with age between five and six are relatively independent of instruction could be tested by a developmental study in a language community such as the Chinese, where the orthographic unit is the word and where reading instruction therefore does not demand the kind of phonemic anlaysis needed in an alphabetic system.

Meanwhile, we are especially concerned to know more about those substantial numbers of first graders, some 30% in our sample, who apparently have not acquired the ability to do phoneme segmentation. It would be of primary interest to know whether they will show deficiencies in reading acquisition as well. We are just beginning this phase of the research. In a recent pilot study, we gave the word-recognition subtest of the Wide Range Achievement Test (the WRAT) to the children who were the first graders of last June's sample. When they are ranked according to their scores on the reading test, we find that while half the children in the lowest third of the class in reading ability had failed the phoneme segmentation test last June, no child in the top third had failed it. Encouraged by these results, we have devised an analytic reading test designed to measure

decoding skills more systematically than is possible with the WRAT. This is now being administered in addition to the WRAT and the phoneme task to a new group of children in Grades 1 and 2.

We have suggested that a lack of awareness of phonemic segmentation may be one serious roadblock to reading acquisition. Data from the analysis of children's reading errors appear to provide additional indirect evidence for this view. It seemed to us that if a child's chief problem in reading is that he cannot make explicit the sound structure of the language, he might be expected to show success with the initial letter which requires no further analysis of the syllable and to show relatively poor performance beyond that point. If all he knows are the letter-to-sound correspondences and that he must proceed from left to right, he might in the case of "bat," for example, simply pronounce the sound for the first letter and then search his lexicon for a word beginning with the sound of that letter. What he needs to do, instead, is to search his lexicon for a word that has three sound segments corresponding to the letter segments in the printed word. However, if he does not know that the words in his lexicon have segments or if he finds phonemic segmentation difficult, he will not be able to map the letters to the segments in those words. By this reasoning, his errors on the final consonants in words should be greater than those on the initial consonants.

We have recently concluded an experiment designed specifically to examine the initial-final consonant error pattern. The subjects were 20 third graders drawn consecutively from the alphabetic registers of a nearby elementary school. The list of words to be read consisted of 38 monosyllables familiar to third graders and selected so as to give equal representation to the 19 consonant phonemes which can occur in both initial and final position in English words. Each phoneme was represented twice in the list in each position. The words were printed on 3 x 5 cards and presented to the child singly to be read aloud to the best of his ability. Testing was carried out in late fall.

Analysis of the data shows final consonant errors to be about twice as frequent as initial (9.5% of the opportunities for final consonants as compared with 4.9% for those in the initial position). A t-test found this difference to be highly significant, with a p value of less than .005. Since it was possible that the difference might be due to the fact that a given phoneme occurring finally may be spelled more complexly than that same phoneme in the initial position (g, j versus dge or ge), we then looked only at the errors on phonemes which are spelled simply (by a single letter) in both initial and final position (p, t, k, b, d, g, m, n, r). If the difference had been due to orthographic complexity, it should have disappeared in this analysis. But it did not. Final consonants still produced significantly more errors (7.8% to 3.0%).

It is clear, then, from these results, that there is indeed a progression of difficulty with the position of the segment in the word, the final consonants being more frequently misread than the initial. Similar findings have been reported by us in a previous study using different word lists (Shankweiler and Liberman, 1972) and by other investigators (Daniels and Diack, 1956; Weber, 1970) who examined error patterns in the reading of connected text. This initial-final consonant difference cannot be accounted for in terms of a simple reflection of the error pattern in speech, as we found in the earlier study of error patterns. There we presented, first for oral repetition and then for reading, a list of -

204 monosyllables chosen to give equal representation to most of the consonants, consonant clusters, and vowels of English. The initial-final consonant error pattern was duplicated in reading, but in oral repetition, the consonant errors were about equally distributed between initial and final position. Moreover, the initial-final error pattern in reading is also contrary to what would be expected in terms of sequential probabilities. If the child at the early stages of beginning to read were using the constraints built into the language, he would make fewer errors at the end than at the beginning of words, not more.

Thus far we have presented several lines of evidence suggesting that the explicit analysis of phoneme segmentation is a hard and unnatural task which may be an important source of difficulty for the child learning to read. But it is certainly not the only serious barrier. The error pattern of vowels provides a case in point. It is well established (Monroe, 1932; Shankweiler and Liberman, 1972; Venezky, 1968; Weber, 1970) that vowels elicit many more errors than consonants. In the segmentation study mentioned above, for example, the vowel errors were twice as frequent as overall consonant errors (15.1% for the vowels as compared with 7.3% for the consonants). It should be noted that this is quite different from what we find in speech. The vowel errors in the oral repetition of speech are infrequent and fewer than those for consonants (Shankweiler and Liberman, 1972).

Why should the error rate for reading vowels be so much higher than that for consonants? It might, of course, be simply because of the embedded medial position of the vowel in the words used to test reading. In order to check on this possibility, we devised a new test consisting of equal numbers of words containing vowels in the initial, medial, and final positions. The seven vowel phonemes that can occur in all three positions were used three times in each position. The words were again monosyllables familiar to third graders. It was found that the overall rate of vowel errors continued to be about twice that of consonant errors (28.3 to 14.0).

There are two reasons at least for suspecting that vowel errors may reflect something other than the segmentation problems which we have suggested as an explanation for the consonant pattern. First, as we have seen, the child can apparently count syllables fairly well and the vowel nucleus stands out in the spoken word as a major element that can be identified in the syllable. A second, and perhaps more interesting reason, comes from a further examination of the error pattern. In the case of consonants, we have noted that errors tend to pile up in the final position. We have taken this as indirect evidence that the child is having segmentation problems. Vowel errors, on the other hand, pattern quite differently. In the third grade study described above, there was no significant difference in error rate for vowels in the initial, medial, and final positions. Moreover, the error rate of vowels in both initial and final position continued to be significantly higher as compared with consonant errors in the corresponding positions (27.6% to 9% in the initial position and 30.5% to 19% in the final position).

Clearly there is no position effect with the vowels; they are simply difficult in all positions. The absence of a position effect may be due to the fact that the vowel is acoustically marked by a burst of sound wherever it appears, while there is no such acoustic mark for the enfolded consonant. In any event, the vowel problem certainly cannot be attributed entirely to segmentation difficulties.

Indeed, we suspect that the errors elicited by consonants and vowels are quite different in their origins. In the case of the consonants, the child has little trouble in learning the spelling-to-sound correspondences. Orthographic complexity makes no appreciable difference to the position effect. The child's error pattern arises mainly from the fact that he cannot map the segmentation of the printed word to the segmentation of the spoken word. The extra difficulties attendant upon the vowels are probably due in part to the obvious orthographic complexities of the spelling-to-sound correspondences, but partly also to the continuous and fluid nature of vowel perception (A. Liberman et al., 1967; A. Liberman, 1970). Though it stands out wherever it occurs in speech, the vowel is complicated by the fact that it can be spelled in many ways in the writing system and is less categorically perceived than the consonants. That is, not only is there a many-to-one mapping of spelling to sound, but because of the continuous nature of vowel perception, even the sound correspondences of single vowel letters may be harder to code and to maintain in memory. We have argued (Shankweiler and Liberman, 1972) that as a consequence of the continuous nature of their perception, vowels tend to be somewhat indefinite as phonologic entities, as illustrated by the major part they play in the variation among dialects and the persistence of allophones within the same geographic locality. By this reasoning, it could be that the noncategorical nature of vowel perception may itself be one cause of the complex orthography and at least one reason why multiple representations of the vowels are tolerated.

The investigation of the effect of orthographic complexity is beset with many problems. To cite only one example: if orthographic complexity is an important source of errors, the number of possible orthographic representations of a given sound should be correlated with the number of errors made on that sound. In fact, however, in a group of second graders we studied recently, the correlation between orthographic complexity and the number of errors lacked statistical significance. Qualitative analysis of the data suggests that this might be due not to the unimportance of orthographic complexity, but rather to the fact that the second grader's knowledge of orthographic rules is so slight that the number of orthographic representations is not yet a relevant factor in determining his errors. We have since developed a cloze procedure test to measure knowledge of orthographic rules for checking our findings, but these data are not yet completed.

Though we believe it is of interest to examine the relation of orthographic complexity of the vowels to the problems of reading acquisition, we recognize that the vowel may be less important in the process than would first appear. It could be argued that if the child's segmentation problems were corrected, his difficulties with the vowels would not be such a serious barrier to reading acquisition. The consonants carry most of the information load. Provided the child knew how many consonants there were and their sequence in the spoken word, an incorrect rendition of the vowel sound would be fairly easily corrected in the context. Surely, getting the vowel correct without a proper analysis of the phonemic structural sequence of the word would be of less benefit to him. If this is so, early teaching methods which emphasize the intensive teaching of the phonemic structure of the word before the introduction of letter forms should be considered. Elkonin (1973), a Russian psychologist, has recently presented considerable experimental evidence that such a method is indeed highly successful.

REFERENCES

Daniels, J. C. and H. Diack. (1956) Progress in Reading. (Nottingham: University of Nottingham Institute of Education).

Doehring, D. G. (1968) Patterns of Impairment in Specific Reading Disability. (Bloomington: Indiana University Press).

Elkonin, D. B. (1973) In Comparative Reading, ed. by J. Downing. (New York: Macmillan).

Fant, C. G. M. (1962) Descriptive analysis of the acoustic aspects of speech. Logos 5, 3-17.

Fletcher, H. (1929) Speech and Hearing. (New York: Van Nostrand).

Gelb, I. J. (1963) A Study of Writing. (Chicago: University of Chicago Press).

Johnson, D. J. and H. R. Myklebust. (1967) Learning Disabilities. (New York: Grune and Stratton).

Klima, E. S. (1972) How alphabets might reflect langauge. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Kolers, P. (1972) Experiments in reading. Sci. Amer. 227 (13), 84-91.

Liberman, A. M. (1970) The grammars of speech and language. Cog. Psychol. 1, 301-323.

Liberman, A. M., F. S. Cooper, D. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.

Liberman, I. Y. (1971) Basic research in speech and lateralization of language: some implications for reading disability. Bull. Orton Soc. 21, 71-87.

Liberman, I. Y., D. Shankweiler, C. Orlando, K. S. Harris, and F. B. Berti. (1971) Letter confusions and reversals of sequence in the beginning reader: implications for Orton's theory of developmental dyslexia. Cortex 7, 127-142.

Makita, K. (1968) Rarity of reading disability in Japanese children. Amer. J. Orthopsychiat. 38 (4), 599-614.

Mathews, M. M. (1966) Teaching to Read. (Chicago: University of Chicago Press).

Mattingly, I. G. (1972) Reading, the linguistic process and linguistic awareness. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambribge, Mass.: MIT Press).

Monroe, M. (1932) Children Who Cannot Read. (Chicago: University of Chicago Press).

Rozin, P., S. Poritsky, and R. Sotsky. (1971) American children with reading problems can easily learn to read English represented by Chinese characters. Science 171, 1264-1267.

Savin, H. (1972) What the child knows about speech when he starts to learn to read. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Shankweiler, D. (1964) Developmental dyslexia: a critique and review of recent evidence. Cortex 1, 53-62.

Shankweiler, D. and I. Y. Liberman. (1972) Misreading: a search for causes. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Stevens, K. N. (1972) Segments, features, and analysis by synthesis. In Language by Ear and by Eye: The Relationships between Speech and Reading, ed. by J. F. Kavanagh and I. G. Mattingly. (Cambridge, Mass.: MIT Press).

Venezky, R. L. (1968) Discussion in Communicating by Language: The Reading Process, ed. by J. F. Kavanagh. (Bethesda, Md.: National Institute of Child Health and Human Development).

Vernon, M. D. (1960) Reading and its Difficulties. (New York: Cambridge University Press).

Weber, R. (1970) A linguistic analysis of first-grade reading errors. Read. Res. Quart. 5, 427-451.

Linguistic and Paralinguistic Interchange*

Philip Lieberman+
Haskins Laboratories, New Haven, Conn.

Current linguistic theory rigidly compartmentalizes the "cognitive," linguistic aspects of human communication and the presumed "emotive," paralinguistic elements that occur in both human and nonhuman communication. The segmental phonetic units of human speech, according to this view, are supposed to convey linguistically relevant information, e.g., the vowel distinction that differentiates the English words bit and bet. Emotive, paralinguistic qualities are supposedly transmitted only by means of prosodic modifications like fundamental frequency, amplitude, and tempo as well as gestures and facial expressions. Nonhuman animals, according to this view, make use only of these paralinguistic parameters. This distinction is false. The same phonetic feature space is used for both paralinguistic and linguistic communication and the semantic boundary line between these two aspects of human communication is not sharp. The foundations of human language can be seen in the paralinguistic aspects of human communication and in the vocal and gestural aspects of the communications of other animals.

Current linguistic theory rigidly compartmentalizes the "linguistic" and "paralinguistic" aspects of human communication. Linguistic communication has been equated to the transmission of cognitive, referential information. Paralinguistic communication has been taken to relate to the transmission of emotive states. Implicit in this distinction is the notion that human language is the medium that allows modern man to think, that, in essence, language is the basis of cognitive ability. Hence the clearly unique aspects of human language, the ability of modern man to form words, phrases, etc., are considered linguistic. In contrast, the prosodic aspects of human language, that is, the modulations of pitch, amplitude, and temporal pattern, which clearly play a part in the communications of other living species besides Homo sapiens, are considered paralinguistic.

---

*For presentation at the IXth International Congress of Anthropological and Ethnological Sciences, Chicago, Ill., September 1973; to be published in the Congress proceedings.

+Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

The distinction between the supposedly paralinguistic and linguistic aspects of communication is misleading. While all animals do make use of innately determined cries to signal certain basic states of their autonomic vegetative systems, no clear distinction can be shown for many of the phenomena that are supposedly paralinguistic or linguistic in human, or for that matter in nonhuman, communication. The gasp of a drowning man is an example of an innately determined cry, as is the cry of a rabbit or dog or a man in extreme pain. Darwin (1872) in The Expression of Emotion in Man and Animals, clearly differentiated these basic cries, which he noted were independent of habit or training, from the emotive information that linguists often classify as paralinguistic. Linguists, in general, tend to classify the transmission of information as paralinguistic when they lack adequate notational systems. If a speaker, for example, told his friend, "The train is due at 8 a.m., but I don't believe it," the information would be treated as a linguistic communication that made use of the speaker's and the listener's cognitive abilities. If the speaker instead had said, "The train is due at 8 a.m.," using a "tone of voice" that conveyed his disbelief, the semantic construct of disbelief would be treated as a paralinguistic phenomenon. Linguists lack adequate transcription systems for the prosodic aspects of language, so they solve the problem by treating as nonlinguistic the information that they cannot describe with their present theoretical and notational apparatus. The situation is ludicrous. It is as though physicists decided that subatomic physics was not part of physics because the present theory could not readily account for the observed phenomena.

We can avoid painting ourselves into this intellectual corner if we consider what we really mean by the term "language." No good analytic definition of language exists. This is an unintended consequence of the search for linguistic "universals": linguists have, for the most part, attempted to define language in terms of the universal properties that structure all human languages. This is an impossible task; we simply do not know what these universals are. If we did, we would have solved the problem of language and would know everything that there is to know about language. The traditional approach towards the definition of language is also anthrocentric in defining language to be necessarily the language of present day Homo sapiens, that is, human language. I think that the following definition of language avoids these problems. I will define a language as a communications system that is capable of transmitting new information. In other words, I am operationally defining language as a communications system that places no restriction on the nature or the quality of the information transferred.

It is obvious that this definition does not require that all languages have all of the properties of human language. It is also obvious that the "phonetic" elements of human language need not be restricted to the segmental phonetic elements that traditional orthography conveys, nor even to the speech signal. Prosodic contours and gestures can have a role even as they do in the languages of other species, living and extinct.

In connection with this last point, it is probable that an advanced hominid species like classic Neanderthal man (Lieberman, 1972), who lacked many of the segmental phonetic elements that characterize human speech, would probably have consistently expressed a semantic construct like disbelief by means of the tone of his voice or a gesture or grimace. The cultural remains of specimens of classic Neanderthal man, Homo sapiens neanderthalis, demonstrate that some form of language must have been present. Fairly abstract cognitive ability must have

been present in these extinct hominids since ritual burials involving the symbolic use of flowers, the use of advanced tools, and the use of fire are all part of the classic Neanderthal culture (Boule and Vallois, 1957; Bordes, 1968; Soleiki, 1972). Present day Homo sapiens has a great segmental phonetic inventory and the semantic construct of disbelief can be expressed either by means of tone of voice or through the use of some additional words.

There clearly is no rigid dichotomy between paralinguistic and linguistic semantic constructs. Any semantic construct that can be paraphrased in terms of a string of words is obviously linguistic. But the use of a phonetic element that cannot be transcribed using the IPA symbol inventory does not make the semantic construct paralinguistic. There is no clear line of demarcation at the semantic level.

No rigid dichotomy exists at the phonetic level with respect to paralinguistic and linguistic phonetic units. A phonetic element is really a signaling unit (Lieberman, 1970). Linguists have been accustomed to manipulating the segmental phonetic elements that are, for the most part, the consequence of the articulatory maneuvers of the supralaryngeal vocal tract in Homo sapiens. Sound contrasts like the vowels [a] and [i], for example, are the result of articulatory maneuvers involving only the supralaryngeal vocal tract (Fant, 1960). Many of the phonetic distinctions that differentiate the segmental phonetic elements are, however, the consequence of laryngeal maneuvers, for example, the distinction between the sounds [b] and [p] (Lisker and Abramson, 1964). The distinction between these two sounds rests in the timing between the start of phonation and the release of the primary occlusion of the supralaryngeal vocal tract. Many languages make use of differences in the dynamic pattern of the fundamental frequency of phonation to signal lexical differences. The various dialects of Chinese, for example, make use of variations in fundamental frequency (which are perceived as pitch variations) to differentiate various words. A speaker of American English does not make use of these distinctions of differentiate the lexical entries of his linguistic "dictionary of words." The speaker of American English is thus free to use these pitch variations, i.e., tone features, to transmit simultaneously the semantic construct of disbelief when he utters the words, "The train is due at 8 a.m." He might also have shrugged his shoulders or used a facial expression that conveys disbelief. The semantic content is nonetheless the same as if he had also added the words, "but I don't believe it."

The speaker thus can make use of phonetic signals that are not intimately associated with the lexical entries in his internal dictionary to convey semantic information that is considered paralinguistic by linguists fixed to the segmental framework of a particular language. In the present company, the particular language in question is English, which many linguists appear to take implicitly as the "universal" language.

There is no clear dichotomy at the phonetic level. Prosodic features that have an exclusive paralinguistic function in one language may have a linguistic, lexical function in another language. High fundamental frequency, rising fundamental frequency, breathy voice, etc., cannot therefore be exclusively viewed as paralinguistic phonetic features. Nor can we even view gestures or facial expressions as exclusively paralinguistic phonetic features. The lexical entries, i.e., "words" of the sign language of the deaf, for example, make use of a wide variety of manual gestures in concert with articulatory maneuvers of the facial

musculature. The language of hominids who lived until comparatively recent times (20,000 - 50,000 years ago), like the people of Shanidar and La Chapelle-aux-Saints, also probably made use of these manual and facial gestures to communicate "words." Present day chimpanzees have, for that matter, been taught to communicate lexical information by means of gestures (Gardner and Gardner, 1969). Chimpanzees exhibit congitive and linguistic abilities that are remarkably similar to, though more limited than, adult modern Homo sapiens (Gardner and Gardner, 1969; Premack, 1972). It is probable that the particular phonetic form of human language is a comparatively recent development in hominid evolution (Lieberman, forthcoming). Cognitive ability, which can take many forms of phonetic expression, must have antedated the appearance of human language.

I do not want to leave the impression that only prosodic and gestural phonetic elements can interchange between conveying linguistic and paralinguistic information. Much of the discussion of the phonetic level of paralinguistic communication is based on either inadequate or incorrect phonetic and acoustic analyses. One often encounters, for example, the assertion that high fundamental frequency conveys some sort of increased emotion on the part of the speaker. Psychoacoustic experiments that transposed the fundamental frequency contour of a synthesized utterance from a "normal" to either "high" or "low" pitch range failed to show this result (Lieberman and Michaels, 1962). The same psychoacoustic experiments demonstrated that the expansion or the compression of the speaker's pitch range also failed to transmit any emotional nuances. These results are in accord with recent acoustic and electromyographic investigations that show great variability in these parameters, both between different speakers and for the same speaker, when completely "unemotional" test sentences are spoken (Atkinson, 1973). The traditional statements concerning the role of pitch that have been constantly repeated and reprinted for at least fifty years are wrong. We simply do not know what is happening.

I must stress that this does not mean that prosodic features do not convey paralinguistic information. The fine structure of fundamental frequency, that is, the variations in periodicity that occur from one opening and closing cycle of the vocal cords to the next, appear to have a paralinguistic function in English (Lieberman, 1961; Lieberman and Michaels, 1962). Dynamic patterns varying the normal prosodic pattern also appear to be relevant. The segmental features also can convey paralinguistic information in English. One of the paralinguistic parameters that speakers normally communicate is their intended sex. (This is not always equivalent to biologically determined sex.) It is obvious that prosodic features convey the speaker's intended sex (Brend, 1971). The segmental phonetic elements also convey the speaker's intended sex. This is obvious in languages that make use of different lexical entries for men and women (Haas, 1964). It is also true in languages like English where speakers use articulatory maneuvers that result in formant frequency differences that distinguish the segmental phonetic elements of men and women (Mattingly, 1966; Schwartz, 1968; Schwartz and Rine, 1968; Sachs, Lieberman, and Erickson, 1972). In effect, men and women have slightly different dialects that involve acoustically and perceptually different vowels and consonants. It also appears that these distinctions are the result of acculturation, that they are learned by children as they learn other aspects of their particular dialect (Sachs et al., 1972). These distinctions in vowel quality, in languages other than American English, can be used to differentiate words. Thus there is a paralinguistic-linguistic interchange, but note that this interchange is again really arbitrary. The speaker's

sex can, if the culture permits, be signaled either through the use of a different word or different set of syntactic or morphophonemic rules, or through the use of a different set of phonetic features. The semantic, cognitive information being transferred is the same; only the means change.

We need not limit our data to the communications of humans, or even primates. The bases of cognitive ability and communication can be seen in the behavior of many species. A dog will signal that he wants water by pushing his water bowl. This is no less an example of cognitive, referential information being communicated than a human requesting a glass of water. We cannot even claim that all of the symbols used by a dog are iconographic. Dogs have been trained to ring bells when they want water. They could not do this unless they had the ability to associate an abstract symbol, the bell, with water. Calling the process "conditioning" does not disguise the cognitive aspects of the problem. Studies of the communications of animals at the neuroelectric level, furthermore, show the presence of "feature" detectors that are tuned to the communicative signals these animals employ (Capranica, 1965; Wollberg and Newman, 1972). The basic principles that structure human communications may be found with the aid of comparative studies of communication in other species. Human language is the result of a long evolutionary process and it involves factors that are important in many aspects of human and animal behavior besides communication (Lieberman, forthcoming). It is immaterial whether the communications are labeled paralinguistic or linguistic; there is no sharp dividing line, for, as Darwin (1859:95) noted, "Natural selection can act only by the preservation and accumulation of infinitesimally small inherited modifications..." Homo sapiens' linguistic abilities only appear to be unique today because the intermediate hominids are extinct. The natural communications of animals like chimpanzee (Goodall, 1971), therefore, are relevant to the study of the basic parameters that underly human language. The paralinguistic-linguistic distinction is again arbitrary.

In conclusion, I think that we should be concerned with the general question of how information is transferred. Whether it is labeled paralinguistic or linguistic is of no concern except to those linguists who want to limit the universe of discourse arbitrarily so that they may claim to have found a "universal" linguistic theory that accounts for all aspects of language.

The principal test of a scientific theory is not that it accounts for everything, but that it relates a number of phenomena that were seemingly unrelated before the theory was proposed. Newton's Laws of Motion never accounted for frictional phenomena. They nonetheless proved correct insofar as they accounted for a diverse range of phenomena that had appeared to be unrelated. To analyze effectively the problem of language, as I have defined language, we have to investigate carefully the acoustic, perceptual, and physiologic parameters that structure language. We have to reexamine many of the premises that are based on either superficial or inadequate analyses and we cannot arbitrarily limit the data sample. We may not be able to account for all of the phenomena that we observe, but we will be in a position to assess both the generality and the limitations of our theories. Only then can we progress.

REFERENCES

Atkinson, J. R. (1973) Aspects of intonation in speech: implications from an experimental study of fundamental frequency. Unpublished Ph.D. dissertation, University of Connecticut.

Bordes, F. (1968) The Old Stone Age. (New York: World University Library, McGraw-Hill).
Boule, M. and H. V. Vallois. (1957) Fossil Men. (New York: Dryden Press).
Brend, R. (1971) Male-female differences in American English intonation. Paper given at 7th International Congress of Phonetic Sciences.
Capranica, R. R. (1965) The Evoked Vocal Response of the Bullfrog. (Cambridge, Mass.: MIT Press).
Darwin, C. (1859) On the Origin of Species, Facsimile edition. (New York: Atheneum, 1967).
Darwin, C. (1872) The Expression of Emotion in Man and Animals. (London: J. Murray).
Fant, G. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).
Gardner, R. A. and B. T. Gardner. (1969) Teaching sign language to a chimpanzee. Science 165, 664-672.
Goodall, J. van L. (1971) In the Shadow of Man. (New York: Dell).
Haas, M. (1964) Men's and women's speech in Koasati. In Language in Culture and Society, ed. by D. Hymes. (New York: Harper and Row).
Lieberman, P. (1961) Perturbations in vocal pitch. J. Acoust. Soc. Amer. 33, 597-603.
Lieberman, P. (1970) Towards a unified phonetic theory. Ling. Inq. 1, 307-322.
Lieberman, P. (1972) The Speech of Primates. (The Hauge: Mouton).
Lieberman, P. (forthcoming) On the evolution of language: a unified view. Proceedings of IXth International Congress of Anthropological and Ethnological Sciences, Chicago, Ill., September 1973. [Also in Haskins Laboratories Status Report on Speech Research SR-33 (this issue).]
Lieberman, P. and S. B. Michaels. (1962) Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. J. Acoust. Soc. Amer. 34, 922-927.
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.
Mattingly, I. G. (1966) Speaker variation and vocal tract size. J. Acoust. Soc. Amer. 39, 1219(A).
Premack, D. (1972) Language in chimpanzee? Science 172, 808-822.
Sachs, J., P. Lieberman, and D. Erickson. (1972) Anatomical and cultural determinants of male and female speech. In Language Attitudes: Current Trends and Prospects, Monograph No. 25. (Washington, D. C.: Georgetown University Monograph Series in Language and Linguistics).
Schwartz, M. F. (1968) Identification of speaker sex from isolated, voiceless fricatives. J. Acoust. Soc. Amer. 43, 1178.
Schwartz, M. F. and H. Rine. (1968) Identification of speaker sex from isolated, whispered vowels. J. Acoust. Soc. Amer. 44, 1736-1737.
Soleiki, R. S. (1972) Shanidar - the First Flower People. (New York: A. Knopf).
Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: response patterns of single cells to species-specific vocalizations. Science 175, 212-214.

Computer Processing of EMG Signals at Haskins Laboratories

Diane Kewley-Port
Haskins Laboratories, New Haven, Conn.

INTRODUCTION

A set of 12 programs for processing electromyographic (EMG) signals has been completed for the Haskins Laboratories' Honeywell DDP-224 computer. Some of the programs and a general description of the EMG hardware and software system appeared in a previous report (Port, 1971). The purpose of this paper is to review the principles underlying the EMG data processing system and to discuss the individual programs in the framework of the tasks they perform.

The primary factor which has determined the structure of the data processing system has been the nature of the EMG signal itself. The electrical signal picked up by a bipolar wire electrode inserted into a muscle is a high frequency signal. However, the speech scientist is primarily interested in the relationship between EMG signals and more slowly varying consequences of the activity, such as muscle tension, force, or the movement of articulators. Of the several methods used to reduce EMG signals, analogue integration is one of the most commonly employed because it has been shown to yield straightforward relationships to other measures of muscle tension, etc. (Inman, Ralston, Saunders, Feinstein, and Wright, 1952; Lippold, 1952; Bigland and Lippold, 1954; Bouisset and Goubel, 1971). Haskins Laboratories has adopted the use of rectified and integrated EMG signals coupled with a 200 Hz sampling rate considered to provide adequate resolution for the investigation of articulatory movement in speech (Cooper, 1965; Gay and Harris, 1971).

An EMG signal picked up by bipolar electrodes has another property of concern to speech scientists known as the interference pattern. The EMG signal observed at the electrodes is the vector sum of biphasic potentials from many motor units such that the waveform from one motor unit can partially cancel the observed potential from another motor unit. Thus, the amplitude of a rectified and integrated EMG signal at any point in time is decreased by the arbitrary phase relationships between motor unit potentials. Since we desire to obtain a quantitative measure of amplitude, we have proposed to remove the phase component by treating it as random noise with a statistical mean of zero. Thus, by taking the average at each point in time of the digitized EMG signals for repetitions of the same utterance, the amplitude on output should correspond to the scaler sum of the motor unit potentials without the phase components which have been summed to zero.

Although the averaging procedure allows us to compare EMG activity quantitatively for a single electrode placement, comparisons of EMG data from separate

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

electrode placements must still be qualitative. Correlation of EMG signals might be a useful technique for determining if the changes in activity observed for one electrode placement vary in a similar way to changes for a different placement. Research is in progress at Haskins Laboratories to determine if the Pearson product moment correlation coefficient is useful in quantifying how similarly two EMG signals vary over time (Port, 1973).

DESIGN OF THE DATA PROCESSING SYSTEM

Based on the principles set out above, we have designed a system to identify many repetitions of a given utterance type and to digitize the corresponding EMG signals, aligning them in time and computing their average. Furthermore, because the system is a tool for speech scientists, care has been taken to ensure that it will have the flexibility and capability to meet their needs.

An EMG experiment conducted at Haskins Laboratories has come to be defined in terms of the fixed storage capabilities of the computer programs. Thus, one EMG experiment can sample up to 8 channels of EMG data simultaneously for up to 30 repetitions of 30 different utterances, each 2 seconds long. The EMG channels, together with an audio channel and a code and timing channel, are recorded on a 16-channel tape recorder. In addition, a calibration signal of 300 microvolts is periodically laid down on each EMG channel.

The step-by-step procedures of conducting an EMG experiment have been described in an earlier report (Port, 1971) and will not be repeated here. Rather, this account focuses on the computer processing of the EMG signals. After the data has been collected on tape, control information for the EMG program must be prepared. This step could be semiautomated on the computer, but currently involves the rather laborious task of visually inspecting the EMG data in the form of oscillographic tracings on Visicorder paper.

During this inspection the experimenter prepares lists of information about each of the utterance types in the experiment. The lists contain the identification of each repetition of an utterance by its unique physical CODE (written on the code and timing channel). Associated with each utterance type is an easily identified event on the audio channel called the LINE-UP POINT, e.g., release of stop consonant. The interval in timing pulses from the CODE to the LINE-UP POINT, called the LINE-UP INTERVAL, is used to align the CODES before averaging. Also, for each utterance a sample window of two seconds or less is specified in timing pulses relative to the LINE-UP POINT. "Housekeeping" information for the whole experiment, such as the muscles investigated and the locations of the 300 microvolt calibration signals, are also gathered. Thus prepared, the experimenter approaches the computer.

COMPUTER AND PROGRAM DESCRIPTIONS

Figure 1, showing a schematic diagram of the computer system, and Table 1, describing the corresponding specifications for each device, are presented as reference information indicating the capabilities of the computer system in use at Haskins Laboratories. The same type of data processing system described here could be implemented on a considerably smaller computer, although it is advantageous to the experimenter to have the large, fast storage capabilities of this system which can rapidly access EMG data across many different experiments.

Figure I

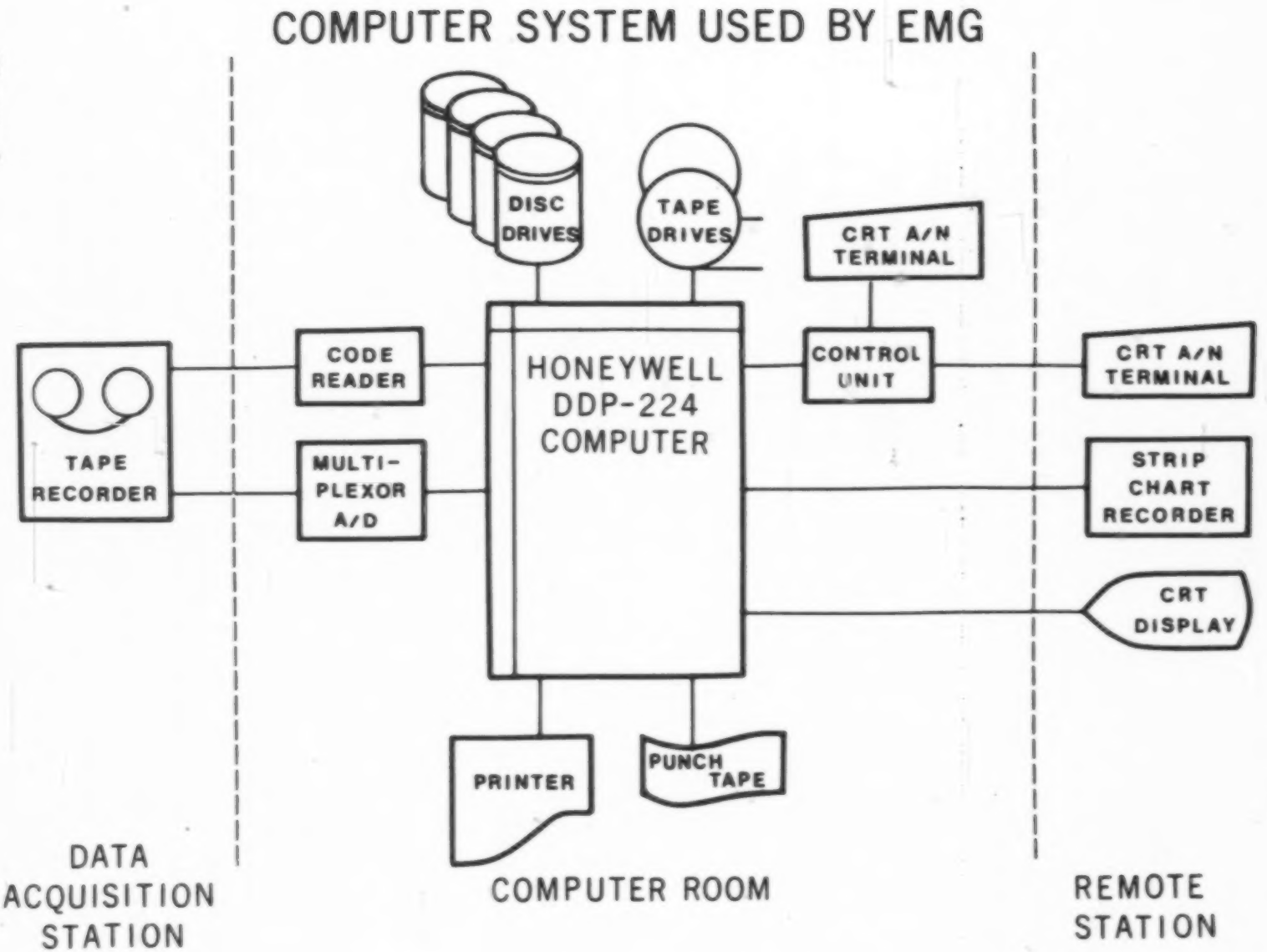

COMPUTER SYSTEM USED BY EMG
DISC DRIVES
TAPE DRIVES
CRT A/N TERMINAL
CODE READER
HONEYWELL DDP-224 COMPUTER
CONTROL UNIT
CRT A/N TERMINAL
TAPE RECORDER
MULTI-PLEXOR A/D
STRIP CHART RECORDER
CRT DISPLAY
PRINTER
PUNCH TAPE
DATA ACQUISITION STATION
COMPUTER ROOM
REMOTE STATION

TABLE 1

HARDWARE SPECIFICATION

DEVICE	SPECIFICATION
COMPUTER	Honeywell DDP-224, 32K memory 24 bits/word, 1.9 μsec cycle time
DISC DRIVES (4)	CDC 9433, moving head 2.4 million words/disc
TAPE DRIVES (2)	CDC 9120, 75 inch/sec. 800 bytes/inch
CRT A/N TERMINALS (3)	Sanders 720 Communicators and control unit
PRINTER	Potter 3502 Chain Printer 132 Char/line, 225 lines/min.
PUNCH TAPE	Berpe 8 channel punch, 110 Char/sec.
STRIP CHART RECORDER	Texas Instruments Rectiriter RRMA, single channel, 12 inch/min.
CRT DISPLAY	Tektronix 611, 12 inch display storage unit
TAPE RECORDER	CE VR-3300 16 channels, FM and AM, 1 inch tape
CODE READER	12 bit buffered octal code reader
MULTIPLEXOR A/D	16 channel Multiplexor, 8 channels used for EMG and input to 12 bit A/D; Multiplexor driven by 3.2K Hz clock track from tape recorder.

The programs have been designed to make the tasks of data sampling and storage, averaging, and hard copy output as automated as possible. Furthermore, several programs are available for the retrieval and manipulation of EMG signals. Ten of the twelve programs are available on our timesharing system, Monitor II. Since Monitor II occupies one of the four disc drives, three disc drives and two tape drives are available for EMG data storage. All the data for a single experiment input or generated by the computer are stored permanently on one digital tape. The programs generally transfer the data from tape to disc for rapid retrieval and storage during processing.

A list of the EMG programs with a brief statement of their purposes appears in Table 2. The programs whose names begin with E$MG are available on Monitor II, the others on Monitor I.

E$MGESEL, ESEL

The first program in the series exists in identical versions on Monitor I (ESEL) and Monitor II (E$MGESEL). The purpose of the program is to create the leading data file on digital tape, containing all the control information for processing EMG data in one experiment. The control information is typed on the alphanumeric terminal. In programming (E$MG) ESEL, emphasis was placed on the ease of man-machine communication. The control information can be printed out for inspection and verification at any time during its operation. An error in any item of control information may be changed at any time.

ECHK

The ECHK program samples the EMG data as specified by the control information and gives feedback information to the operator for editing. For example, one task is to adjust amplifier gains on playback so that the level for each EMG channel is near the maximum sample range of seven bits without overloading. This is done by presenting on-line on the alphanumeric terminal the maximum and minimum values after sampling each CODE. (In the program descriptions, "CODE" will signify both the octal number and the corresponding data sampled for each repetition of an utterance.) The maxima and minima are also stored in the form lists of CODES for each utterance and are printed out at the end of a pass of ECHK.

Gross errors in the specification of CODES or LINE-UP INTERVALS in the control information can be detected by scanning the ECHK print-out of maxima and minima for consistency of the data in each channel. Corrections in the control information must be made using ESEL, but it is an easy matter to switch back and forth between ECHK and ESEL.

ERIT

All the EMG data specified in the control information are sampled and stored on disc during one pass of ERIT. Twelve bits of data are sampled every 5 msec using 8 channels of a 16-channel multiplexor and one analogue-to-digital converter. The data are transferred from discs to digital tape when sampling is completed.

TABLE 2

EMG COMPUTER PROGRAMS

PURPOSE	PROGRAM
SPECIFY UTTERANCE TYPES CODES AND LINE-UP INTERVALS	E$MGESEL, ESEL
SET LEVELS, GROSS EDITING	ECHK
SAMPLE AND STORE	ERIT
COMPUTE SUMS, SUMS OF SQUARES	E$MGSUMS
VISUAL EDITING ON CRT	E$MGPAGE
PRINT AVERAGES, STD. DEV.	E$MGSUMS
PLOT HARD COPY, TEST PLOTTER	E$MGPLOT, E$TSTPLT
EXAMINE AVERAGE EMG SIGNALS ON CRT	E$MGDISP
CORRELATION OF EMG SIGNALS FOR ONE EXPERIMENT	E$MGCOR1
CORRELATION OF EMG AVERAGES FOR SEVERAL EXPERIMENTS	E$MGCOR2
HARDWARE SYSTEMS TEST	E$MGTEST
PACKED PAPER-TAPE OUTPUT	E$MGTAPE

E$MGSUMS

The two tasks performed by E$MGSUMS are, by historical accident, now separated in the data processing sequence by E$MGPAGE. The first task is to compute the sums and sums of squares at 5 msec intervals over the sample window for each utterance. Only 7 of the 12 bits of data sampled are significant and are used in this computation. From the samples of the 300 microvolt calibration signals a conversion factor for each channel is calculated. The sums and sums of squares and conversion factors are stored on the digital tape for final visual editing using E$MGPAGE. When the editing is completed, E$MGSUMS is used again to calculate and print out the means and standard deviations in microvolts for each 5 msec interval. For an experiment of 30 utterances, each one second in duration, the printout takes about 45 minutes for about 200 full pages (11 x 14 inches). The sheer bulk of this data has prompted the development of computer programs to help the experimenter manipulate and display the averaged EMG signals on-line.

E$MGPAGE

Even a careful experimenter has trouble correcting all faulty control information using ESEL and ECHK, thereby permitting some unrepresentative CODES to be averaged in E$MGSUMS. E$MGPAGE has therefore been written to enable the experimenter to scan rapidly through one or more EMG channels for an entire experiment, comparing the individual repetitions of an utterance with the corresponding average for that utterance. Figure 2 shows a picture of the CRT display. The experimenter is usually looking for two types of errors: an EMG signal may appear shifted in time relative to the LINE-UP POINT for some CODE, or there may be nonphysiological spikes embedded in an EMG signal. Whatever the problem, the experimenter indicates on the display the number of the CODE to be removed, and the program automatically deletes it from the list of CODES for that utterance, recalculates the sums and sums of squares, and displays the new EMG average with the appropriate CODES for further inspection. When the experimenter is satisfied that the sums and sums of squares contain only representative CODES, E$MGPAGE automatically constructs a corrected digital tape. This tape is used for printing the averages in E$MGSUMS.

The example of the display produced by E$MGPAGE, shown in Figure 2, illustrates the usefulness and reasonableness of the averaging technique in eliminating the interference pattern. The high-frequency jitter of the interference pattern is clearly seen for the CODES (numbers 9 to 14) on Figure 2. In contrast, the average EMG signal is smooth. Moreover, important information such as the onset or offset of EMG activity, and relative peak heights seems well preserved in the average signal. A detailed examination of individual differences in utterance repetitions is currently under investigation (Port, 1973).

E$MGPLOT, E$TSTPLT

The strip chart recorder can be used to produce individual graphs of the average EMG signals. E$TSTPLT produces a test pattern to calibrate the recorder. E$MGPLOT outputs selected EMG signals where the y-axis maximum is specified separately for each channel in microvolts. The paper speed, and therefore data rate, on the recorder is quite slow so the program is generally timeshared. We expect to phase out this plotter soon and replace it with a Tektronix 613 display scope which will have a hard copy unit (model 4610) attached to it for reproducing displays from the E$MGDISP program.

POLAROID PHOTOGRAPH OF CRT DISPLAY FROM E$MGPAGE

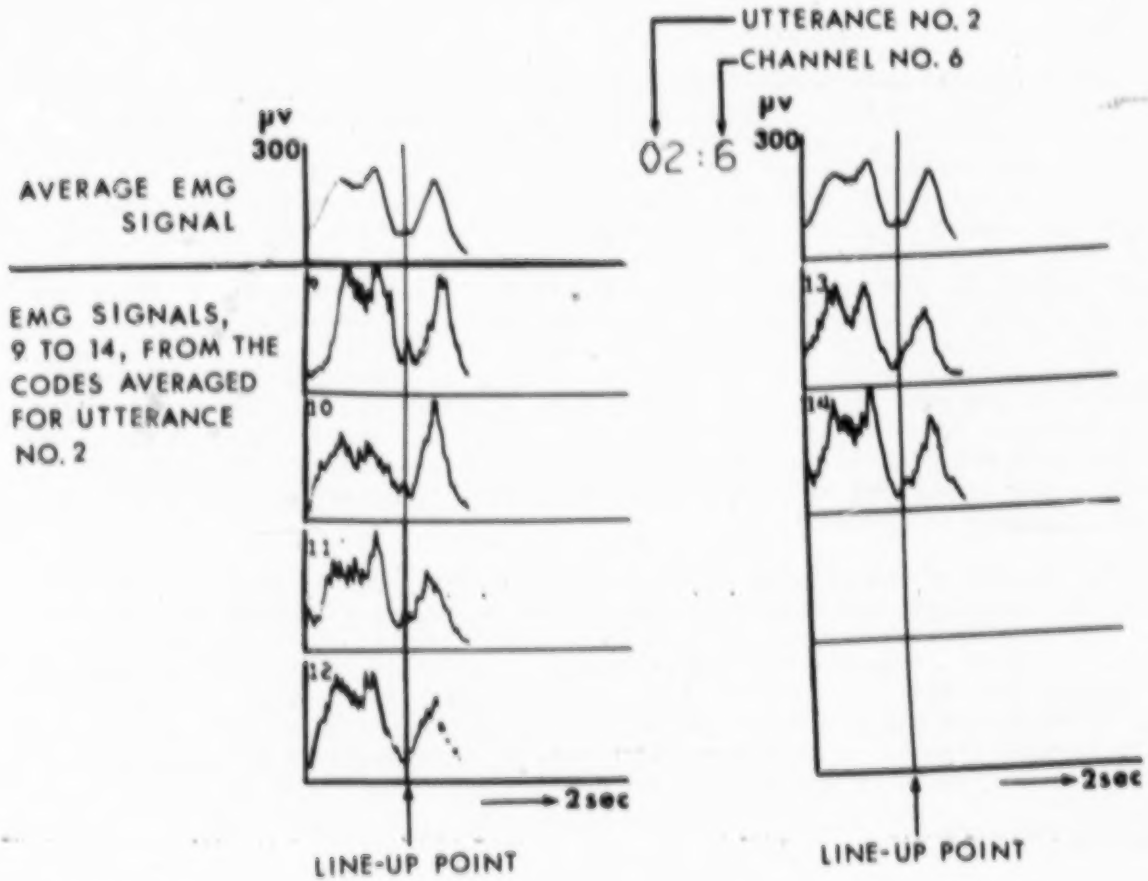


Figure 2

E$MGDISP

The task of retrieving selected EMG signals for comparison on the CRT display is done by E$MGDISP. A photograph of the display is found in Figure 3. The experimenter can compare up to nine EMG signals on the display at one time. There are three line types displayed for each of three sets of axes. Number codes to the left of each set of axes are referenced to a print-out from E$MGDISP and indicate which EMG signals have been displayed. Currently, an experimenter can have available on disc all the EMG signals for up to 12 different experiments (coded by tape number). E$MGDISP, then, is the main tool the experimenter uses for comparing EMG data, and for assisting him in evaluating and forming his hypotheses. A single-shot Polaroid camera fitted with a hood for the CRT display has been used frequently to take photographs from E$MGDISP for reference and for slides or figures for manuscripts.

E$MGCOR1, E$MGCOR2

These two programs are used to compute a Pearson product moment correlation coefficient, r, for EMG signals. The experimenter has the option of having r computed over the entire sample window for two specified EMG signals, or of indicating which subset of the samples is to be correlated.

Using E$MGCOR1 the correlation coefficient may be calculated between any two EMG signals obtained from a single experiment. For example, r may be calculated for any two CODES or between any CODE and its corresponding average EMG signal. E$MGCOR2 is used to compute r for average EMG signals obtained from up to 12 different EMG experiments available on disc. The printer is used for output.

E$MGTEST

E$MGTEST is used to check that the hardware in the playback system is operating correctly. A standarized steady-state signal is recorded on the tape recorder prior to each EMG experiment for each of the 8 EMG channels. E$MGTEST is used to sample these signals continuously and display the digitized values on the alphanumeric terminal. The operator can then check that the values are indeed consistent and within the expected limits. The program is run only periodically or for troubleshooting.

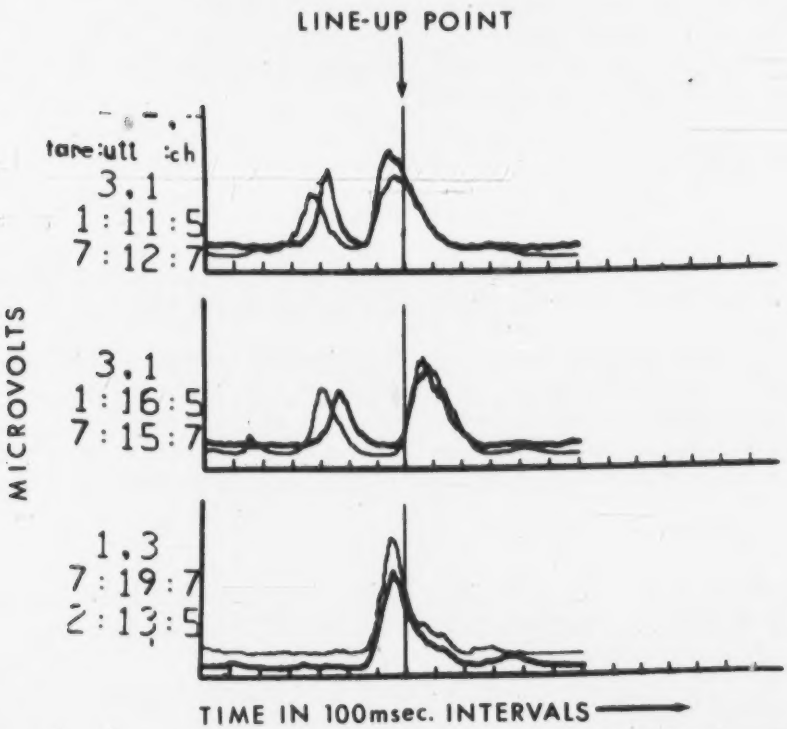
E$MGTAPE

Packed paper-tape output of the sums of the EMG signals (before averaging) is available for visitors to Haskins Laboratories who wish to input EMG data to a different computer facility. One experiment can be recorded on about 400 feet of 8-channel paper tape. The tape can be checked for parity errors.

FUTURE PROGRAMMING

No major additions to the EMG computer programming system are anticipated in the near future. However, a new computer facility is expected to be installed at Haskins Laboratories early in 1974 and it will eventually be necessary to transfer the EMG programs. The new programs will probably bear a close resemblance to the current ones, which are performing satisfactorily and have an established role as a tool for speech research at Haskins Laboratories.

POLAROID PHOTOGRAPH OF CRT DISPLAY FROM E$MGDISP



CODE FOR AXIS PAIR	EXAMPLE (TOP AXIS) (THIRD LINE--NOT PRESENT)
Y-AXIS MAXIMA IN 100 MICROVOLTS	3, 1, (, NONE)
LINE TYPE	
— TAPE NO.: UTT. NO.: CHAN. NO.	1: 11: 5
▬ TAPE NO.: UTT. NO.: CHAN. NO.	7: 12: 7
-- TAPE NO.: UTT. NO.: CHAN. NO.	(NONE)

Figure 3

REFERENCES

Bigland, B. and O. C. J. Lippold. (1954) The relation between force, velocity, and integrated electrical activity in human muscles. J. Physiol. 123, 214-224.

Bouisset, S. and F. Goubel. (1971) Interdependence of relations between integrated EMG and diverse biomechanical quantities in normal voluntary movements. Activitas Nervosa Superior 13, 23-31.

Cooper, F. S. (1965) Research techniques and instrumentation: EMG. In Proceedings of the Conference: Communicative Problems in Cleft Palate, ASHA Reports No. 1, 153-168.

Gay, T. and K. S. Harris. (1971) Some recent developments in the use of electromyography in speech research. J. Speech Hearing Res. 14, 241-246.

Inman, V. T., H. J. Ralston, J. B. de C. M. Saunders, B. Feinstein, and G. J. Wright. (1952) Relation of human electromyogram to muscular tension. EEG Clin. Neurophysiol. 4, 187-194.

Lippold, O. C. J. (1952) The relation between integrated action potentials in a human muscle and its isometric tension. J. Physiol. 117, 492-499.

Port, D. K. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

Port, D. K. (1973) Techniques for processing EMG signals. (abstract) J. Acoust. Soc. Amer. 53, 320.

Pitch Determination by Adaptive Autocorrelation Method

Georgije Lukatela[+]

ABSTRACT

This paper describes an adaptive autocorrelation method for pitch determination which has several useful features: 1) the capability of tracking rapid pitch excursions despite the low amplitudes common to many voiced consonants; 2) the capability of analyzing the whole range of pitch variation in human speech; 3) immunity to a moderate degree of degradation of speech quality; and 4) feasibility of real time operation, with a constant delay of 50 msec. A software implementation of this method is discussed and evaluated.

INTRODUCTION

During the last 40 years, and especially during the last decade, a large variety of methods for determination of the fundamental frequency of pitch of speech have been proposed. Excellent reviews of these methods are available (McKinney, 1965; Schroeder, 1966). Earlier work in pitch tracking was related mostly to general speech analysis and vocoder applications. Recently, there has been considerable interest also in pitch determination in connection with speech recognition and speaker identification problems. Over the years, pitch determination methods have evolved from simple low-pass filtering to rather sophisticated processing of the speech signal. Some of the methods have proved to be very successful in determining pitch, but the most reliable method has still been visual inspection of the speech waveform. Obviously this laborious method is not suitable for any large-scale statistical measurements.

This paper describes an improved autocorrelation method for determining the duration of pitch periods. In its software version, the method is suitable for general linguistic research using a small computer, and in its hardware version it offers a real-time pitch extractor that is not unreasonably complex.

[+]University of Beograd, Yugoslavia. The research reported here was done while the author was a Fulbright Fellow at Haskins Laboratories, 1969-70.

Acknowledgment: I am indebted to Dr. Franklin S. Cooper for many valuable discussions and for his encouragement, to David Zeichner and Richard Music for their friendly assistance, and to Diane Kewley-Port for her willingness to write the instructions.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

AUTOCORRELATION ANALYSIS PROBLEM

As pointed out by Schroeder (1966) a prerequisite for the successful solution of the pitch determination problem in speech has been the separation of the spectral envelope from the spectral fine structure. This separation has been carried out by cepstral methods and, indeed, short-time cepstrum analysis has worked particularly well in pitch detection. Unfortunately, the hardware implementation of cepstral methods is laborious; moreover, these methods can fail when the voiced part of the speech spectrum consists of only a few spectral components.

The study reported here was prompted by Sondhi's work (1967) on the autocorrelation of spectrum-flattened speech, and undertook to develop further the autocorrelation method, taking advantage of some previously unexploited properties of speech signals. As is well known, the autocorrelation function (τ) of a periodic signal $s(t) = s(t + T_o)$ has a maximum for zero delay $(\tau = 0)$, for a delay of one pitch period $(\tau = T_o)$, for a delay of two pitch periods $(\tau = 2T_o)$, and so on. These maxima are of the same height and shape. Although the speech signal is not periodic, over short time intervals it can be considered quasiperiodic. Hence, the short-time autocorrelation function has been defined (Schroeder and Atal, 1962) as:

$$\phi(\tau,t) \equiv \int_{-\infty}^{t} f(x) \cdot f(x - |\tau|) \cdot \gamma_{\phi}(t - x)dx \qquad (1)$$

where $\gamma_{\phi}(t)$ is a physically realizable but otherwise arbitrary "weighting function." $\gamma_{\phi}(t)$ can be visualized as a "window" through which a part of the input signal is viewed. For this reason, $\gamma_{\phi}(t)$ is also called a "lag window."

In the general case, the short-time autocorrelation function of the speech signal may be taken as the sum of three components:

$$\phi(\tau,t) = \phi_{PP}(\tau,t) + \phi_{FP}(\tau,t) + \phi_{AP}(\tau,t) \qquad (2)$$

The first component, ϕ_{PP}, takes account of pitch periodicities; the second component, ϕ_{FP}, of formant (basically first-formant) periodicities; and the third component, ϕ_{AP}, reflects the presence of aperiodicities. Assuming that the width of the lag window is comparable to the pitch period, one may observe some typical shapes of the "ϕ-pattern," i.e., of the short-time autocorrelation function of speech as plotted against the delay τ. Over the quasiperiodic intervals of speech, when the pitch periodicity is dominant, the ϕ-pattern shows an absolute maximum for zero delay $(\tau = 0)$ and several maxima at $\tau = kT_o (k = 1, 2, 3, \ldots)$. The heights of the subsequent maxima are usually less than that for zero delay. Incidentally, it may be noted that $\phi(0,t)$ represents the mean square value of the speech signal $s(t)$.

For $\tau > 0$, the aperiodic component ϕ_{AP} in (2) vanishes rapidly, and the formant component ϕ_{FP} behaves like a damped oscillation. However, during fast pitch inflections or rapid formant glides, the pitch periodicity term may become less than that for the formant periodicity, i.e., the pitch peaks in the ϕ-pattern can show multiple false peaks, and the tracking of the "pitch peak" that corresponds to the true pitch period may become a difficult task. A possible solution of this problem, so fundamental to the autocorrelation analysis, is discussed in the next section.

PRINCIPLES OF THE PROPOSED METHOD

In studying the distribution and relative values of peaks in the ϕ-pattern of low-pass filtered speech, it was noticed that the voiced segments of speech start with a rapid build-up of the $\phi(0,t)$ value and that the peak at $\tau = T_o$ is well defined, provided the width of the lag window is large. Since this initial build-up behavior of voiced speech segments has shown small variance across many speech samples, it was designated as an invariant feature of the speech signal. During the Initial-Build-Up (IBU) period, it was possible to locate the true maximum of the short-time autocorrelation function (for $\tau = T$) even if the inflection is fast because the speech signal energy increases rapidly, and the previous maximum is easily updated by the subsequent one.

Thus, by the end of the IBU period, there is reliable value for the true pitch peak, which makes it possible to follow closely the subsequent movements of this peak. This involves narrowing the lag window to the minimum width consistent with accumulation of the necessary speech signal energy. This narrow lag window allows updating of the pitch peak position even during the rapidly decreasing speech levels usually found at the end of phonation. Tracking of the pitch peak during the post-IBU period is safeguarded by several additional precautionary measures. These consist of comparison of the new pitch period value (T_o) with the previous one and comparison of the assumed $\phi(T_o,t)$ value with the weighted values of other peaks in the ϕ-pattern.

When the input speech signal is band limited (telephone quality), preprocessing is mandatory. In an early study by Licklider and Pollack (1948), it was shown that center-clipping the speech signal drastically reduces its intelligibility. Many years later, Sondhi (1967) demonstrated that this center-clipping provides a very efficient and simple means of formant removal. This suggested further that any nonlinear transformation of the speech waveform that emphasizes the highest instantaneous peaks is appropriate for speech spectrum flattening. For instance, raising the speech signal to a higher odd power (third, fifth, etc.) transforms the speech waveform into a quasiperiodic train of impulses having a broad, almost uniform power spectrum up to the frequency.

$$f_{un} = \frac{1}{2\theta} \tag{3}$$

where θ is the effective width of impulses. However, the width of the impulses can be made too narrow, thereby substantially reducing the intensity of the spectral components, and resulting in a poor signal-to-noise ratio at the autocorrelator input. As a compromise, raising the speech signal to the third power seemed to give the best over-all performance. Results with a center-clipping method, using a self-adjusting bias as the reference level for the clipping operation, were somewhat inferior.

OUTLINE OF THE SYSTEM

A block diagram of a system for pitch determination which was implemented at Haskins Laboratories is shown in Figure 1. It consists basically of a preprocessor, correlator, maxima selector, postprocessor, and control unit. The preprocessor carries out the functions of automatic volume control and nonlinear transformation that are necessary for formant removal. When the input speech is of high quality, the preprocessor can be omitted. The autocorrelator generates a spatial set of short-time autocorrelation functions, i.e., it transmits to the maxima selector a spatial ϕ-pattern. This component finds and stores

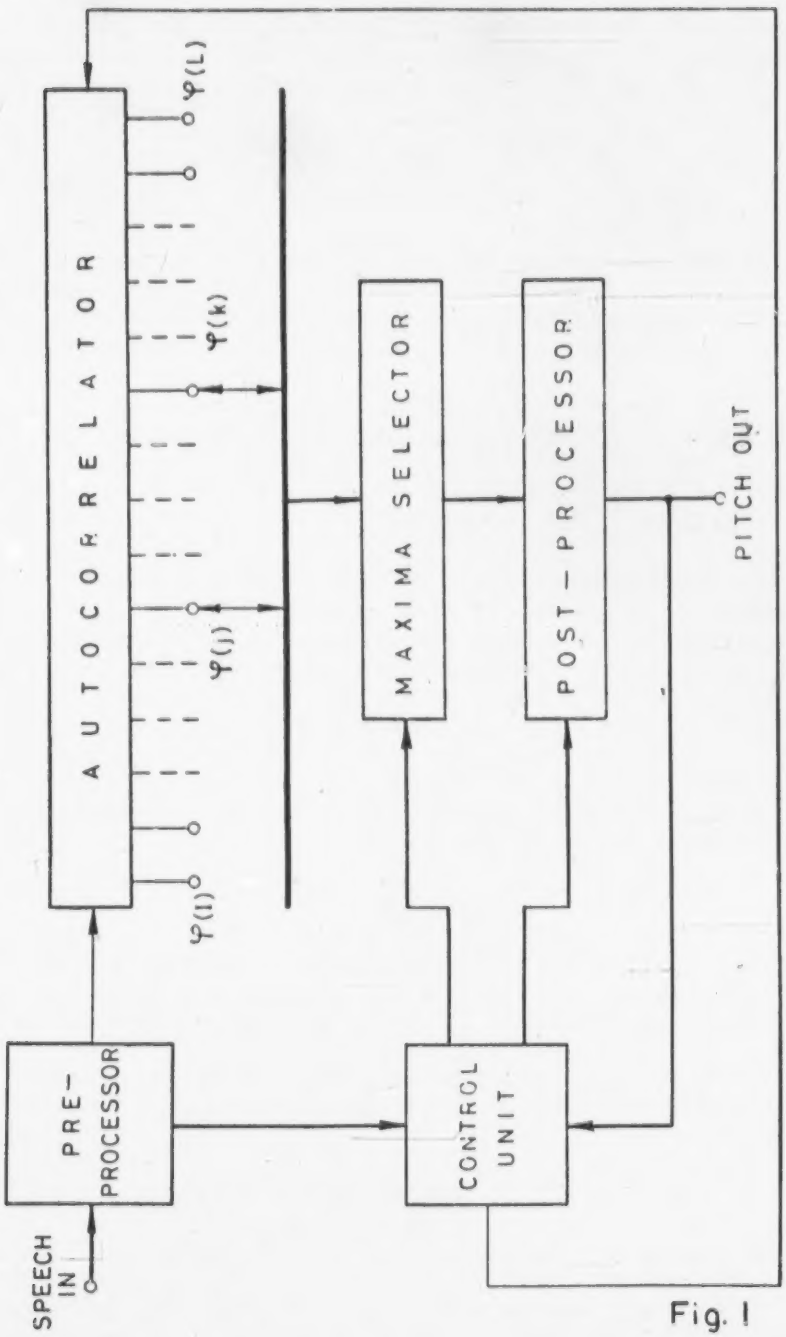

SPEECH IN
PRE-PROCESSOR
AUTOCORRELATOR
φ(1)
φ(j)
φ(k)
φ(L)
MAXIMA SELECTOR
POST-PROCESSOR
CONTROL UNIT
PITCH OUT

Fig. 1

a number of maxima, while the postprocessor checks and compares the heights and delays of the stored maxima and determines which maximum represents the instantaneous value of the pitch period. With this information, and considering also the mean square value of the input signal, the control unit makes appropriate changes in the time constants of the autocorrelator and in the position and width of the "spatial window" through which the maxima selector views the spatial ϕ-pattern.

The autocorrelator consists of a band-pass filter, an amplitude sampler, a delay line with N output taps, and a set of N multipliers and integrators. The band-limited speech signal, sampled at a rate of 8 kHz, propagates down the delay line. At each output tap, a short-time autocorrelation function is formed. For the sake of simple implementation, the definition of the short-time autocorrelation function as given by Fano (1950) was used:

$$\phi(\tau, t) \equiv \frac{1}{\gamma} \int_{-\infty}^{t} f(x) \cdot f(x - \tau)\, e^{-\frac{t - x}{\gamma}}\, dx \qquad (4)$$

The integration is carried out by means of a simple RC integrator with a time constant γ. During the IBU period (which lasts 50 msec), the value of the time constant is γ = 14 msec. In the post-IBU period this value is automatically decreased to 8 msec. For exceptionally low-pitched male speakers, a somewhat larger value of γ = 9 msec is set.

SYSTEM PERFORMANCE

In order to evaluate the over-all performance of the proposed method, an autocorrelation program was written for the Haskins Laboratories DDP-224 computer which operated on speech waveforms reduced to digital form. Analyses were made of various difficult phrases and sentences pronounced by adult male and female speakers and by children. A few baby cries were analyzed also. The utterances were recorded and their waveforms and sonograms were analyzed by visual inspection. In this way the "true" pitch contours were determined.

Representative pitch diagrams obtained by the adaptive autocorrelation method are shown in Figures 2 and 3. The instantaneous fundamental frequency (F_o)--the reciprocal of the pitch period duration (T_o)--is plotted against time. The sentence in Figure 2, "Palm trees grow very tall," was spoken by a low-pitched male voice. This utterance is interesting because of its rapid and relatively large pitch changes. The pitch contour in the voiced fricative [v] in the word "very" was determined despite a fast pitch change at low power level. In the last word of this sentence ("tall"), the pitch lowers within a 300 msec interval from about 140 Hz to about 35 Hz, i.e., about two octaves. The whole contour in the voiced alveolar [ℓ] is faithfully reproduced. The pitch diagram in Figure 3 was derived from an emphatically spoken question, "Is it pattering?", asked by a female speaker. The first and second words were uttered briefly and at a very low power level. The last syllable ("-ing") is remarkable for the extreme pitch inflections used to signal emphatic stress.

Figures 2 and 3 show an interesting feature: at the onset of voicing following each voiceless stop there is a high falling pitch. This contrasts with what occurred at the onset of voicing after voiced stops, namely, either a low rising or a steady pitch. These observations support findings recently reported (Haggard, Ambler, and Callow, 1970).

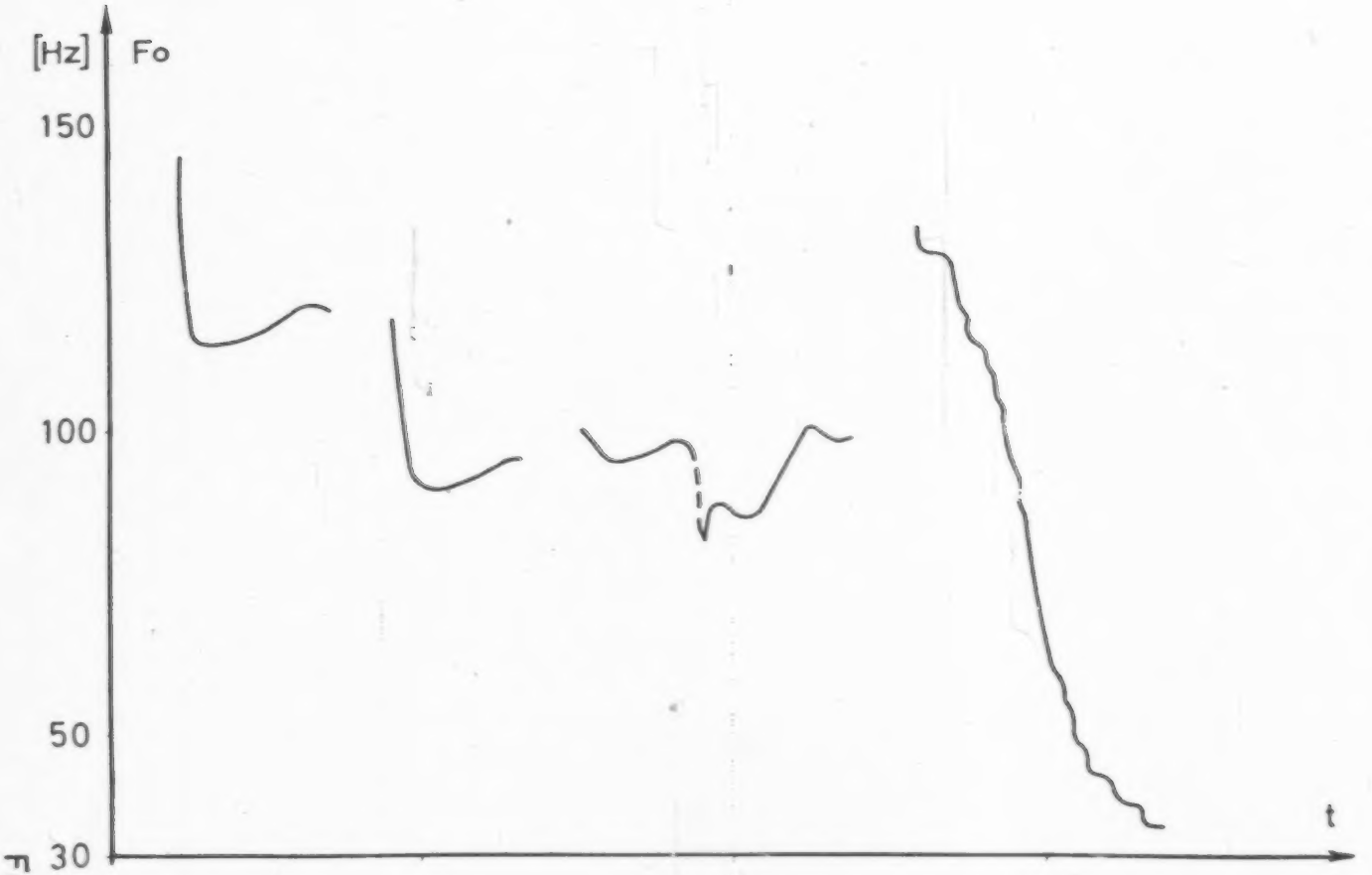
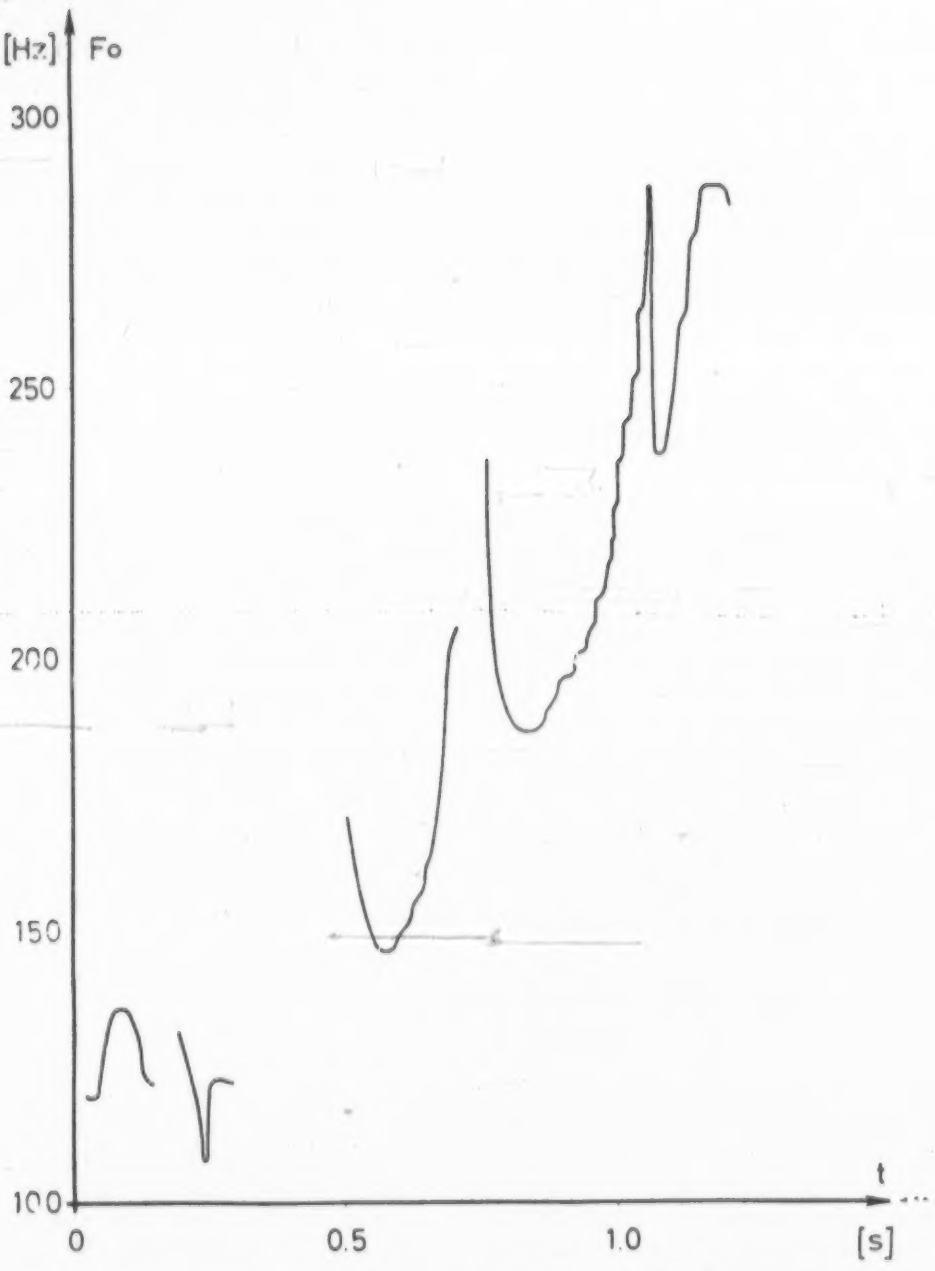


Fig. 2

P A LM TR E E S GROWVERY T A L L



ISI TP ATTERING

Fig. 3

In a series of measurements, the relation between the input speech quality and the system performance was analyzed. It was possible to obtain basic pitch information from speech signals with additive white noise in an approximately zero signal-to-noise condition. In those cases, with an average spectral level of white noise about 10 dB or more below the speech spectral components (measured within the frequency band of the first formant) there are only minor disturbances in pitch determination. However, if the applied noise contains periodic components that are only slightly lower than the high-pass cut-off frequency of the input filter, a serious operational impairment can occur: the system is likely to lock up on the interfering periodic noise component.

The high-pass filtering of the input speech is a matter for concern. If the high-pass cut-off frequency does not exceed approximately 100 Hz, the pitch determination is not seriously deteriorated, as compared with full-band speech. Moreover, for this condition, preprocessing of the input speech is not obligatory. For example, the pitch diagrams in Figures 2 and 3 were obtained with no preprocessing and with a high-pass cut-off frequency of about 90 Hz. However, when the frequencies below 250 Hz have been removed from the input speech, preprocessing becomes very important as a means of neutralizing the adverse effects of the high-pass filtering. By comparison, low-pass filtering of the input speech is completely irrelevant unless the low-pass cut-off frequency is decreased below about 1200 Hz.

In this method, there are two possible kinds of errors: regular and nonregular. The regular error is a small deviation from the true value. It arises either as a result of the quantization in the time measurement or because of system inertia. Nonregular error, on the other hand, represents a serious failure in pitch determination. The experimental results so far obtained show that the probability of a nonregular error is very low. In about 300 voiced segments of connected speech that have been analyzed, only two nonregular errors were found: 1) a voiced segment was completely omitted because its duration was short and its pitch change was too fast, and 2) the pitch period duration of a voiced segment was halved because, during the IBU period the wrong maximum in the ϕ-pattern was selected.

The regular quantization error is due to granularity in the measurement of time. With an 8 kHz sampling rate the elementary delay for each cell of the delay line is $\tau_o = 0.125$ msec. Hence, the maximum quantization error in the determination of fundamental frequency is given by:

$$\frac{\Delta F}{F_o} = \pm \frac{1}{2} \frac{\tau_o}{T_o + \frac{1}{2}\tau_o} \approx \pm \frac{1}{2} \frac{\tau_o}{T_o} \qquad (5)$$

Thus, at the highest frequencies of interest (about 500 Hz) the relative error rate would amount to $\pm$ 3 percent, whereas for very low frequencies (about 40 Hz) it would be only $\pm$ 0.25 percent. During moderate pitch variations at normal signal levels the accuracy of pitch determination has been limited by quantization errors only. However, when the pitch change is fast and the signal level is low, the maximum total error may amount to about $\pm$ 10 percent of the nominal instantaneous fundamental frequency.

CONCLUSION

The adaptive autocorrelation method has shown the possibility of determining, with small probability of failure, the pitch of connected speech, particularly during rapid pitch inflections and transitions. The method can work over the whole range of pitch variation encountered in human speech, i.e., from 30 Hz to 600 Hz. These features of the proposed method could be useful for general linguistic research.

REFERENCES

Fano, R. M. (1950) Short-time autocorrelation functions and power spectra. J. Acoust. Soc. Amer. 22, 546-550.

Haggard, M., S. Ambler, and M. Callow. (1970) Pitch as a voicing cue. J. Acoust. Soc. Amer. 47, 613-617.

Licklider, J. C. R. and I. Pollack. (1948) Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. J. Acoust. Soc. Amer. 20, 42-50.

McKinney, N. P. (1965) Laryngeal frequency analysis for linguistic research. Communication Sciences Lab., University of Michigan, Ann Arbor, Report 14 (September).

Schroeder, M. R. (1966) Vocoders: analysis and synthesis of speech. Proceedings IEEE 54, 720-734.

Schroeder, M. R. and B. S. Atal. (1962) Generalized short-time power spectra and autocorrelation functions. J. Acoust. Soc. Amer. 34, 1679-1683.

Sondhi, M. M. (1967) New methods of pitch extraction. 1967 Conference on Speech Communication and Processing, Conference Preprints (Cambridge, Mass.: MIT) 239-243.

An Electromyographic Study of the American English Liquids*

David R. Leidner

With the advent of the sound spectrograph, phoneticians and linguists had before them a means of objectifying the acoustic qualities of speech sounds. In 1952 Jakobson, Fant, and Halle attempted to describe the minimal functional units of speech in acoustic terms. Phonologists found these acoustic terms quite useful in formulating phonological rules. As phonological research progressed, however, it was found that certain data could be accounted for only by a gymnastic use of the feature set because the articulatory aspects of the production of sounds were ignored in the formulation of phonological features. Recognizing this, Chomsky and Halle (1968) suggested a feature inventory based on articulatory configurations of the vocal tract. This had the mechanical advantage of capturing naturalness and of explicitly showing feature interaction. But Chomsky and Halle only occasionally discuss the acoustic and perceptual correlates of the features, not because the features are uninteresting or unimportant, but because, they say, "...such a discussion would make this section...much too long" (p. 299). One can reasonably infer from this that the acoustic qualities of sounds play little, if any role in their distinctiveness. A further conclusion would be that a fully satisfactory explanation of phonological phenomena--both synchronic and diachronic--can be made solely by reference to the way sounds are articulated. Finally, assigning a unique articulatory configuration to each phone assumes that there is a one-to-one mapping between the articulatory feature set assigned to the sound and the acoustic output.

---

*Paper presented at the 47th annual meeting of the Linguistic Society of America, Atlanta Ga., December 1972.

+The author is a student of Dr. Arthur S. Abramson, Haskins Laboratories and Department of Linguistics, University of Connecticut. The experiments discussed here were carried out at Haskins Laboratories, and partial support for the research was provided by a Dissertation-Year Fellowship from the University of Connecticut Research Foundation.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

The American English liquids are given the following feature set by Chomsky and Halle (1968):

	/r/	/l/
voc	+	+
cons	+	+
high	-	-
back	-	-
low	-	-
ant	-	+
cor	+	+

This is not meant to account for the allophonic variations in liquids: presumably, low-level variations would be taken care of by low-level phonetic output rules. The primary purpose of this study was to examine and explain the various allophones of the liquids by studying the activity of some of the muscles involved in their production. Specifically, we are interested in the activity of the lips and tongue in the production of American English /l/ and /r/ intervocalically, word-finally, and preconsonantally. Bipolar wire electrodes were inserted into the lips and tongue muscles of two subjects who read lists of /l/ and /r/ in various contexts 19 times. The raw data were stored on FM tape, inspected visually for artifacts, and averaged by computer to get a general picture of the contribution of each muscle. Some selected aspects of the experiments will be discussed.

Figures 1 and 2 show orbicularis oris superior, styloglossus, and tongue tip activity for one subject producing the set ir-iri-irp-irb-irm (Figure 1) and il-ili-ilp-ilb-ilm (Figure 2). The line drawn in the center of each curve represents the onset of phonation of the first vowel of the test utterance, with 400 msec to the left of the line and 600 msec to the right. Each utterance token was read in the frame "It's a h___" to provide for a neutral phonetic environment. Notice that for both /r/ and /l/ before the labials /p/, /b/, and /m/ the orbicularis oris superior shows two peaks after the line-up point. The second peak corresponds to the closing gesture for the labial consonant, an observation supported by high-speed cinematographic data taken during the experimental runs. The first peak cannot be associated with lip spreading for /i/, since it occurs considerably after the onset of phonation of the vowel--generally, this first peak occurs at least 100 msec after the onset of phonation, when the lips have already relaxed their spreading gesture.

Turning now to the activity of the styloglossus, we can see that for both /l/ and /r/, styloglossus activity is much less for the intervocalic member than for preconsonantal liquids: the styloglossus muscle, which can pull the tongue upward and back, is thus in part responsible for what is commonly called "velarization." Correlated with this in the articulation of /l/, though not of /r/, is a lessened tongue tip activity for the nonintervocalic set. Notice that

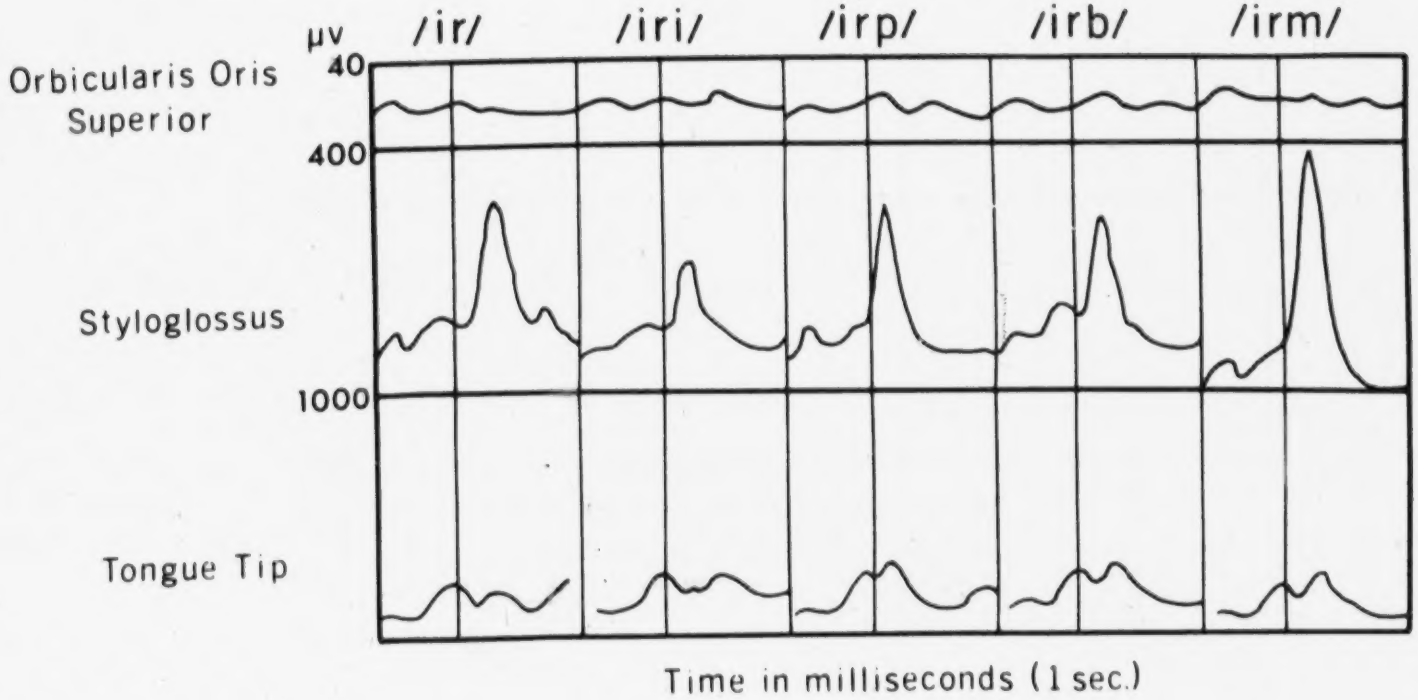

/ir/
/iri/
/irp/
/irb/
/irm/
μv
40
400
1000
Orbicularis Oris Superior
Styloglossus
Tongue Tip
Time in milliseconds (1 sec.)

Figure 1

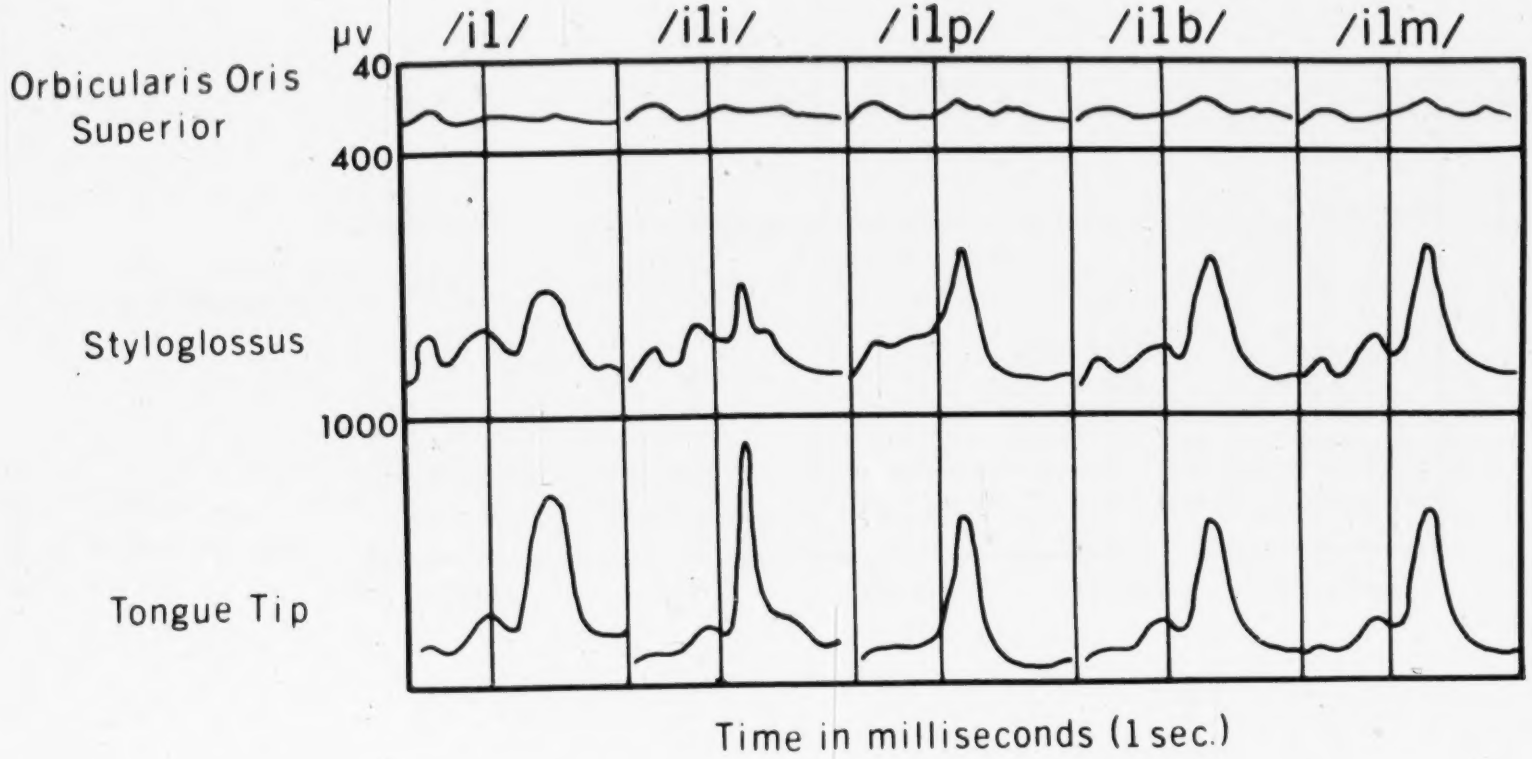

μv
/i1/
/i1i/
/i1p/
/i1b/
/i1m/
Orbicularis Oris Superior
40
400
Styloglossus
1000
Tongue Tip
Time in milliseconds (1 sec.)

Figure 2

where there is less tongue tip activity there is a corresponding increase in styloglossus activity. We might speculate that tongue tip activity decreases either because the liquid is in word-final position--a phenomenon not uncommon for final consonants in languages generally--or because of a pressure to articulate the next consonant--in other words, a tendency toward consonant-cluster simplification. In any event, the significant observation is that the corresponding increase in styloglossus activity--i.e., velarization--cannot be explained on mechanical grounds: there is no articulatory factor which forces the styloglossus to act when tongue tip activity falls. Similarly, no mechanical explanation can be found for the extra orbicularis oris superior peak for preconsonantal liquids.

Data for the second subject, not provided here, yield similar results: the upper lip shows two peaks, both more closely associated temporally than in the first subject. For /l/, tongue tip activity is greatest when in intervocalic position, while for /r/, the anterior gènioglossus, which can pull the tongue upward and forward, has greatest activity for the intervocalic position and least activity for /ir/ and for the prelabial set.

When the vowel is /a/ instead of /i/, slightly different data are obtained, as would be expected, but the effect is similar. Here, the posterior portion of the medial intrinsic tongue muscles, which upon contraction pull the tongue backward and possibly upward as well, show greater activity for preconsonantal /l/, while tongue tip activity is greatest in the intervocalic /ala/. For /r/ in the /a/ series, the styloglossus again shows least activity for the intervocalic member. We thus again find a tendency for velarization where tongue tip activity falls, while there are also two peaks in upper lip activity. Roughly similar results are obtained when the vowel is /u/, but the picture is much too complex to discuss here. Suffice it to say that here, too, a prelabial liquid calls for greater velarization than with the intervocalic liquids.

The interpretation of these data is enhanced when viewed in light of the behavior of liquids across languages. Consider, for example, the extra peak for the upper lip. If this is a rounding gesture, as I suggest, then we would expect coarticulatory effects to show up in the area of phonology. One example is a Viennese dialect of German, in which front vowels round before /l/ (Trubetzkoy, 1969:232). In certain nonstandard Czech dialects, a velarized /l/ can both round and back a preceding vowel (Entwistle and Morison, 1964:320). In Brazilian Portuguese, word-final alveolar /l/ often becomes /U/ (von Essen, 1964). In French, we find an /l/ ~ /o/ alternation in masculine nouns which end in /l/ in the singular, but in /o/ in the plural (Schane, 1968:51-52). Some Slavic languages show this change (Entwistle and Morison, 1964). In English, the Cockney dialect shows this quite clearly, as do certain midwestern and southern dialects of the United States, though not as pronounced. The phoneme written as /r/ varies widely in articulation, but it too is subject to velarization. For example, the earliest phase of breaking in Old Norse--the splitting of a vowel into a rising diphthong whose second member is a back vowel--occurred only when /e/ was followed by /r/ + consonant or /l/ + consonant (Flom, 1937). Similarly, /r/ or /l/ + consonant prevented /a/ > /e/--i.e., fronting--in Old Saxon and Old Frisian (Prokosch, 1939:116-117).

Putting all these observations together, we can clearly see that: (1) liquids tend to be "velarized" preconsonantally and, to an extent, word-finally; and (2) a

lip gesture often accompanies the loss of the liquid. The facts of language correspond quite closely with the observed EMG data described above. Let us consider, now, a hypothetical and much simplified case of both /l/ and /r/ becoming /w/. This would be handled under the framework presented by Chomsky and Halle (1968) as:

$$\begin{bmatrix} +\text{ voc} \\ +\text{ cons} \end{bmatrix} \longrightarrow \begin{bmatrix} -\text{ cons} \\ -\text{ voc} \\ +\text{ back} \end{bmatrix}$$

Since we do not have to specify the precise articulation of the liquids, we can let the rule apply to the many variations of liquids found in languages. After applying the marking conventions for glides to the second term, we would get a fully specified matrix for /w/. What is bothersome about this approach is not only that a rule of this sort fails to explain the change in the fullest sense of the term, but also that the change seems to be an unnatural one. It would be simpler, for example, in a feature-counting sense, for the liquids to become either a true vowel or a true consonant: in other words, in a change in either [consonant] or [vocalic]. Looked at another way, we find phonemes--/l/ and /r/--whose allophonic variations--/u/ or /w/--are not related in the same sense as are, say, the advanced and retracted varieties of /k/, or a consonant with its palatalized counterpart, for example /k/ and /c/. In other words, we cannot account for the extra lip-protrusion gesture on these mechanical grounds.

The only way to explain the particular substitution is by reference to the acoustic properties of the liquids. Data provided in such studies as Lisker (1957), O'Connor, Gerstman, Liberman, Delattre, and Cooper (1957), and Lehiste (1964) show that the liquids are similar in formant structure to /w/ or /u/ in the first three formants: /r/, like /w/, has a low F_2, while /l/, like /w/, has a relatively high F_3 as well as a low F_2. Furthermore, both liquids show an especially low second formant when in word-final position. One would expect this to happen in preconsonantal position. These similarities are brought into higher relief when /y/ is considered.

We can now begin to explain the observed velarization and the tendency toward lip protrusion. Under certain circumstances--namely, those conducive to lessened tongue-tip articulation--the speaker seeks to maintain the acoustic quality of the liquid phoneme by using compensatory maneuvers. In the case of the liquids, there are two ways to doing this, both of whose aim is to increase the area in front of the major tongue constriction: one can hump the tongue posteriorily or one can lengthen the vocal tract by protruding the lips, or both. Either maneuver results in a lowered F_2, which, among other factors, characterizes the liquids acoustically.

This explanation falls in line with Lieberman's (1970) Unified Phonetic Theory, which argues that not only are there preferred articulatory configurations for a sound, based on "acoustic stability" factors, but that different articulatory maneuvers can be used to achieve the same acoustic effect. This implies that in the case of certain phonemes--i.e., distinctive feature bundles--the acoustic consequences form a target with preferred articulatory configurations. Due to such factors as sluggishness of the articulators or a following semi-antagonistic gesture involving the same muscle, the speaker uses an entirely

different gesture to preserve the acoustic identification of a phoneme. And it is this acoustic goal which must be included in the distinctive feature matrix of certain phonemes in order to account for certain gestures found, at first, on the level of phonetic implementation.

We can also thus explain the observed pronunciation of /l/ and /r/ in the early speech of children. Since they cannot see how the tongue is positioned, they try to reproduce the acoustic output of the adult model. Hence, their rendering of "wabbit" for "rabbit" and "sweepy" for "sleepy." No doubt this principle operates not only with the liquids but with other sounds as well. Such a proposal has in fact been put forth by Menyuk and Anderson (1969). These observations have further consequences for what has been termed the "motor theory of speech perception." Rather than claiming that speech sounds are perceived in the way they are produced, we would amend this to say that speech sounds are perceived in the way or ways they could have been produced.

REFERENCES

Chomsky, N. and M. Halle. (1968) The Sound Pattern of English. (New York: Harper and Row).

Entwistle, W. J. and W. A. Morison. (1964) Russian and Slavonic Languages. (London: Faber and Faber).

Essen, O. von. (1964) An acoustic explanation of the sound shift [ɬ] > [u] and [l] > [i]. In In Honour of Daniel Jones, ed. by Abercrombie et al. (London: Longmans) 53-58.

Flom, G. T. (1937) Breaking in Old Norse and Old English. Language 13, 123-136.

Jakobson, R., C. G. M. Fant, and M. Halle (1952) Preliminaries to Speech Analysis. (Cambridge: MIT Press).

Lehiste, Ilse. (1964) Acoustical Characteristics of Selected English Consonants. (Bloomington: Indiana University Research Center in Folklore, Anthropology, and Linguistics) Publication 34. (Also numbered as Internat. J. Amer. Ling. 30, 3.)

Lieberman, P. (1970) Toward a unified phonetic theory. Ling. Inq. 1, 307-322.

Lisker, Leigh. (1957) Minimal cues for separating /w, r, l, y/ in intervocalic position. Word 13, 256-267.

Menyuk, P. and S. Anderson. (1969) Children's identification and reproduction of /w/, /r/, and /l/. J. Speech Hear. Res. 12, 39-52.

O'Connor, J. D., L. J. Gerstman, A. M. Liberman, P. C. Delattre, and F. S. Cooper. (1957) Acoustic cues for the perception of initial /w, j, r, l/ in English. Word 13, 24-43.

Prokosch, E. (1939) A Comparative Germanic Grammar. (Philadelphia: Linguistic Society of America).

Schane, S. A. (1968) French Phonology and Morphology. (Cambridge: MIT Press).

Trubetzkoy, N. S. (1969) Principles of Phonology, tr. by Christiane A. M. Baltaxe. (Berkeley and Los Angeles: University of California Press). (Originally published as Grundzüge der Phonologie. Prague: Travaux du Cercle Linguistique de Prague, No. 7, 1939.)

The Role of the Extrinsic and Intrinsic Tongue Muscles in Differentiating the English Tense-Lax Vowel Pairs*

Lawrence J. Raphael[+] and Fredericka Bell-Berti[++]
Haskins Laboratories, New Haven, Conn.

Phoneticians and phonologists generally agree in recognizing three pairs of vowels on the basis of the proximity of the members of each pair to each other in the vowel space. These pairs, shown in Figure 1, are /i/ and /I/, both high front vowels; /e/ and /ɛ/, both mid front vowels; and /u/ and /U/, both high back vowels. There is far less agreement, however, as to what feature constitutes the essential difference between the members of each pair. Among the differentia proposed are: tongue tension, in which the usually unspecified muscles of the tongue are said to be more tense for /i e u/ than for /I ɛ U/; duration, in which /i e u/ are said to be long vowels whereas /I ɛ U/ are short; diphthongization, in which /i e u/ are said to be characterized by quality or color changes, whereas /I ɛ U/ are simple, monophthongal vowels; tongue height, in which /i e u/ are simply said to be articulated with slightly higher tongue positions than /I ɛ U/; and jaw opening, in which /i e u/ are said to be articulated in a more close jaw position than /I ɛ U/.

While any theorist may recognize the presence of most, if not all, of these features in the articulation of the vowel pairs, it is possible to select one feature as the essential one and to demonstrate logically that the other features are redundant or automatic reflexes of it.

Among the differentia already mentioned, duration, diphthongization, tongue height, and jaw opening can all be studied more or less directly by electromyographic procedures. These features are all, however, observable by other techniques as well. It is the tense-lax opposition which can be investigated by electromyography as by no other means. The EMG curve, in terms of both its peak height and the total area beneath it gives a highly reliable measure of muscular activity.

In studying the feature of tension, it was our aim to discover which, if any, of the muscles of the tongue were consistently more tense in the production

---

*Paper presented at the American Speech and Hearing Association Convention, San Francisco, Calif., November 1972.

[+]Also Herbert H. Lehman College of the City University of New York.

[++]Also the Graduate School of the City University of New York, and Montclair State College, Upper Montclair, N. J.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

Fig. 1

of the purportedly tense vowels than in the production of the other, purportedly lax vowels. It was our expectation that, because of their often antagonistic functions, not all of the extrinsic and intrinsic tongue muscles would exhibit a tense-lax opposition in the frequently hypothesized direction; e.g., /i/ as tense, /I/ as lax.

The utterances used in this experiment consisted of the six vowels already mentioned, produced in a consonant-vowel-consonant (CVC) syllable preceded by schwa. The syllable-initial consonant was always /p/, as was the final consonant. Between 15 and 20 utterances of each type were averaged to produce the EMG curves. The EMG data were obtained from hooked-wire electrodes inserted into the various muscles by means of a hypodermic needle.

Studies on three extrinsic and three intrinsic muscles are reported here. The extrinsic muscles are the genioglossus, the styloglossus, and the palatoglossus. The genioglossus has been shown to be active in raising the bulk of the tongue, particularly for the articulation of front vowels. The styloglossus has been thought to "draw the tongue upward and backward" (Zemlin, 1968). The palatoglossus has long been assumed to raise the back of the tongue when the velum is fixed (Zemlin, 1968; Kaplan, 1971), although a better description of its function might be that it serves to narrow the opening to the oropharynx by approximating the anterior faucal pillars.

The first of the intrinsic muscle fibers we shall discuss here are those of the inferior longitudinal muscle, whose function is presumed to be the shortening of the tongue and/or the pulling down of the tongue tip (Zemlin, 1968; Palmer, 1972). The second electrode placement into the body of the tongue cannot be specified as to particular muscle fibers. The insertion was made in the anterior portion of the tongue near the midline, through the superior surface, an area where the unsheathed fibers of the intrinsic tongue muscles mingle freely. Based on current anatomical information, we assumed that most of the fibers contributing to the EMG signal from this electrode placement would be transverse or vertical. For want of a better name, we have called the source of the EMG signal from this placement the "central fibers."

The superior longitudinal was the last of the intrinsic tongue muscles investigated. It showed no significant activity for either subject.[1] In other experiments, some performed on the same day as those reported here and with the same electrode placements, we have observed activity in this muscle, especially for tongue-tip consonants. We thus tend toward the conclusion that the muscle is not active in vowel articulation, although further confirmation should be sought. In any event, we shall not comment in detail on the action of this muscle.

We shall look at the extrinsic tongue muscles first. The genioglossus, again, is active in raising the bulk of the tongue. The data (Figure 2) for this muscle, which we have previously reported (Raphael, 1971a, 1971b), indicate that for both subjects there was greater activity for /i/ than for /I/, for /e/ than for /ɛ/, and for /u/ than for /U/. (The zero point in time in this and all

---

[1] In a preliminary study of Japanese vowels, done at Haskins Laboratories by Hirose and Hiki (personal communication), no activity was found for the superior longitudinal.

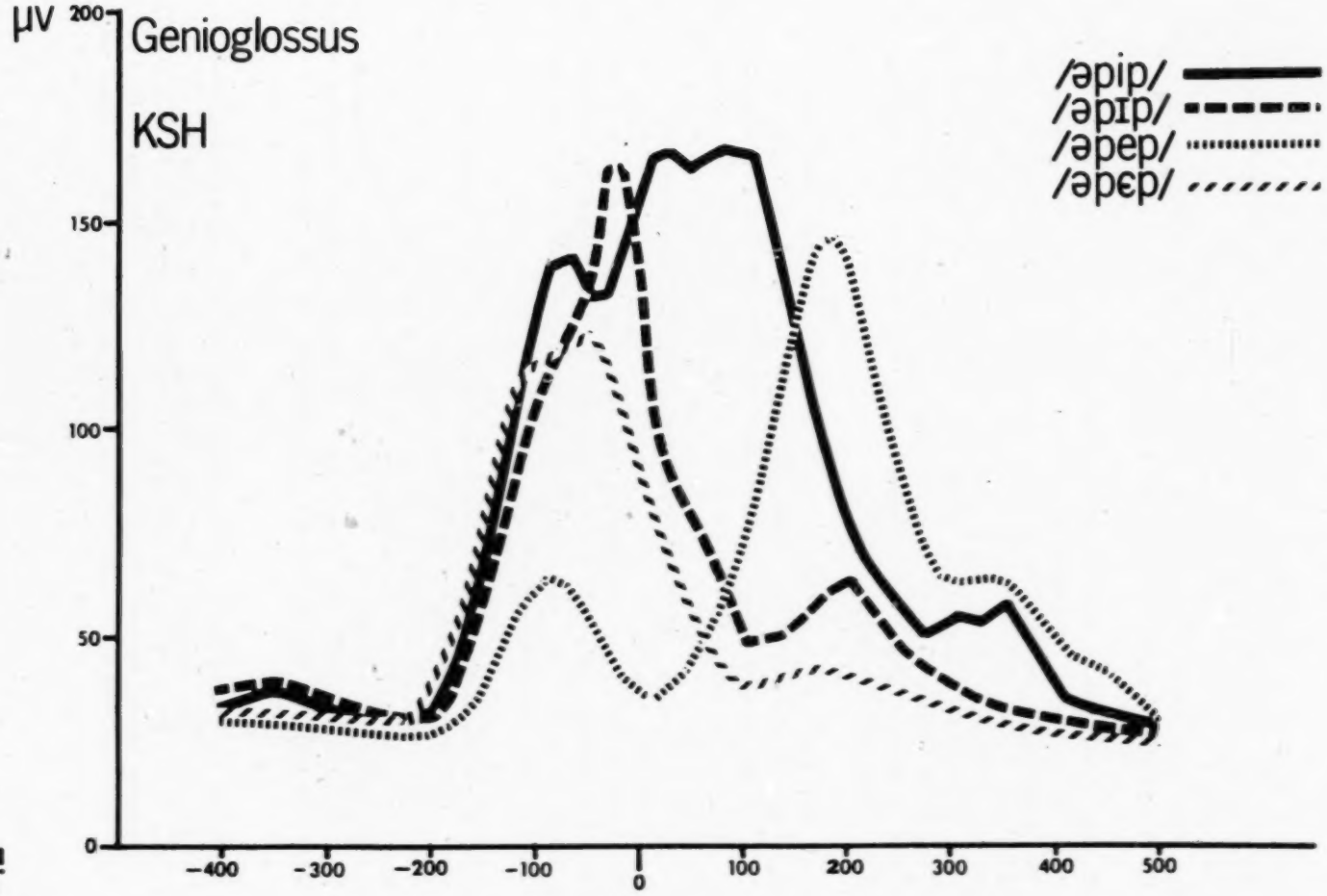

μv
200
150
100
50
0
Genioglossus
KSH
/əpip/
/əpɪp/
/əpep/
/əpɛp/
−400
−300
−200
−100
0
100
200
300
400
500
time in msec.

Fig. 2

other figures is the onset of the voicing of the vowel.) For subject number one the front tense vowels also exhibit greater duration and diphthongization, the latter shown by the bimodal nature of the EMG curve. The back vowels for this subject (Figure 3) both show less EMG activity than the front vowels, as expected. The durational difference is again evident, but the diphthongization is less pronounced for /u/ than it is for /i/ and /e/.

For the front vowels (Figure 4), subject number two shows much the same picture in terms of tension and durational differences, although there is no clear evidence of diphthongization. The same may be said of the back vowels (Figure 5) for this subject. That is, both the peak and the duration of the EMG curve are greater for /u/ than for /U/, but the curve for /u/ is unimodal.

Thus, the genioglossus muscle evidences a consistent tense-lax difference between the vowels /i e u/ and /I ɛ U/, although it must be pointed out that such differences cannot be shown to be independent of differences in duration and change in vowel color.

The styloglossus muscle, which is assumed to draw the tongue upward and backward, shows quite a different pattern from the genioglossus. Figure 6 shows the /e/-/ɛ/ pair which is quite representative of the styloglossus activity for subject number two. In general, there was little or no difference in activity between the members of the vowel pairs (here the scale is 100 mv), or, as in Figure 7, an example from subject number one, there was slightly more activity on the part of the vowel usually termed "lax."

The only exception to this was the second subject's /i/-/I/ pair, in which /i/ showed appreciably greater activity than /I/ in the styloglossus. There seems to be no regular difference in duration and no direct indication of diphthongization in the EMG traces for either subject. On the basis of these results we can establish no consistent tense-lax difference between vowels for the styloglossus muscle.

Data for the palatoglossus muscle (Figure 8), which has been assumed to be active in raising the back of the tongue, were obtained only for subject number two. As expected, the palatoglossus showed significant activity only for the back vowels. For this subject greater activity is found for /U/, the vowel usually referred to as lax, than for /u/, the vowel usually labelled tense. The data for this subject and for this vowel pair thus indicate a reversal of the tense-lax opposition as traditionally conceived. They also, obviously, raise questions about the usually assumed function of the palatoglossus muscle mentioned above.

The inferior longitudinal muscle (Figures 9 and 10), on the other hand, displays just the type of tense-lax differences usually hypothesized. This muscle is thought to be active in depressing the tongue tip and shortening the body of the tongue. For subject number two there is consistently and obviously greater muscular activity for /i e u/ than for /I ɛ U/. The data show greater activity for the front vowels (and slightly more for /i/ than for /e/). The back vowels (Figure 11) show less activity than the front vowels. In all, the inferior longitudinal data for this subject show a striking resemblance to the genioglossus data, with gross differences in total activity and duration between the curves for the tense and lax vowels, but with no direct evidence of diphthongization. Similarly, the inferior longitudinal and genioglossus data for

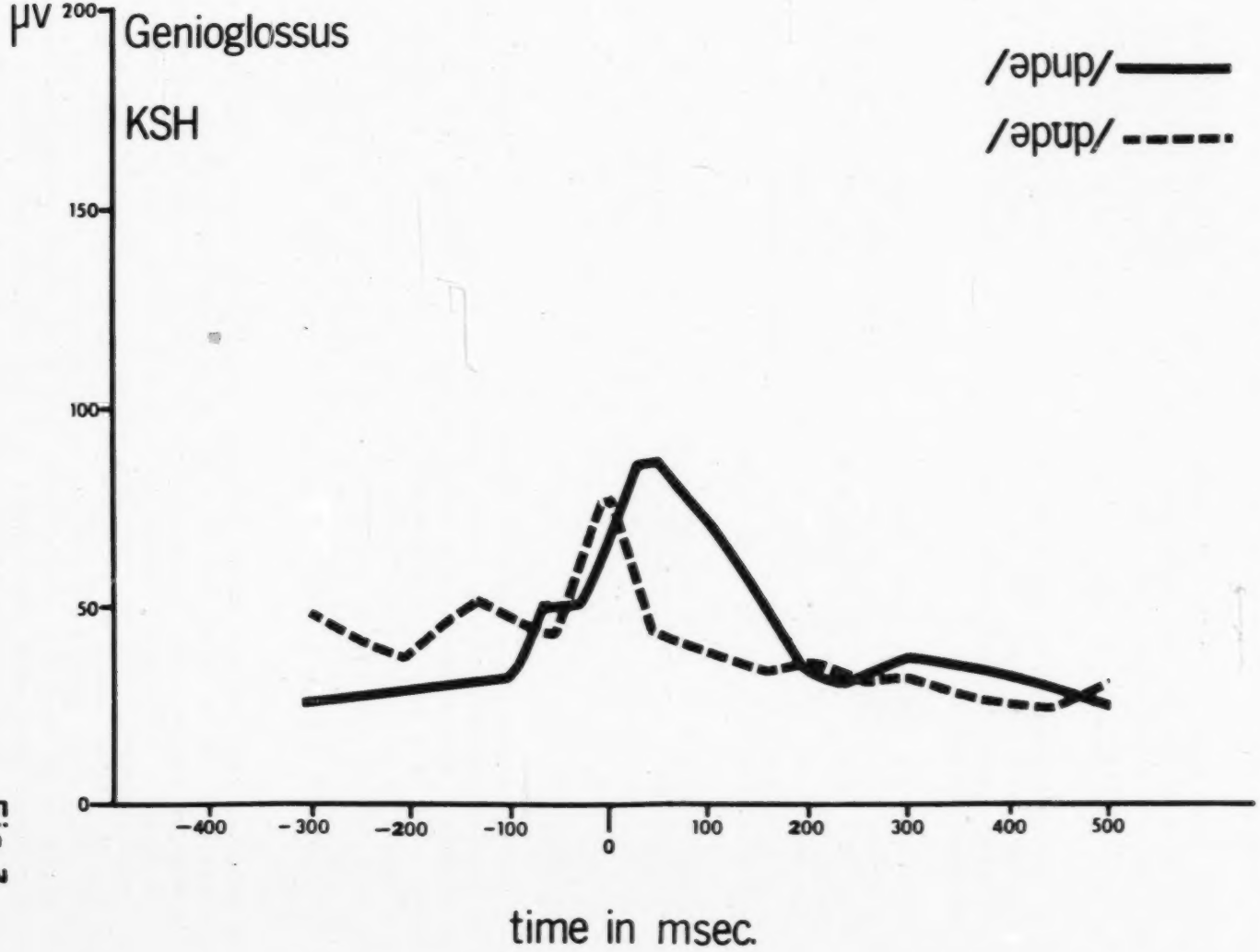

μv 200
Genioglossus
KSH
/əpup/
/əpup/
150
100
50
0
−400
−300
−200
−100
0
100
200
300
400
500
time in msec.

Fig. 3

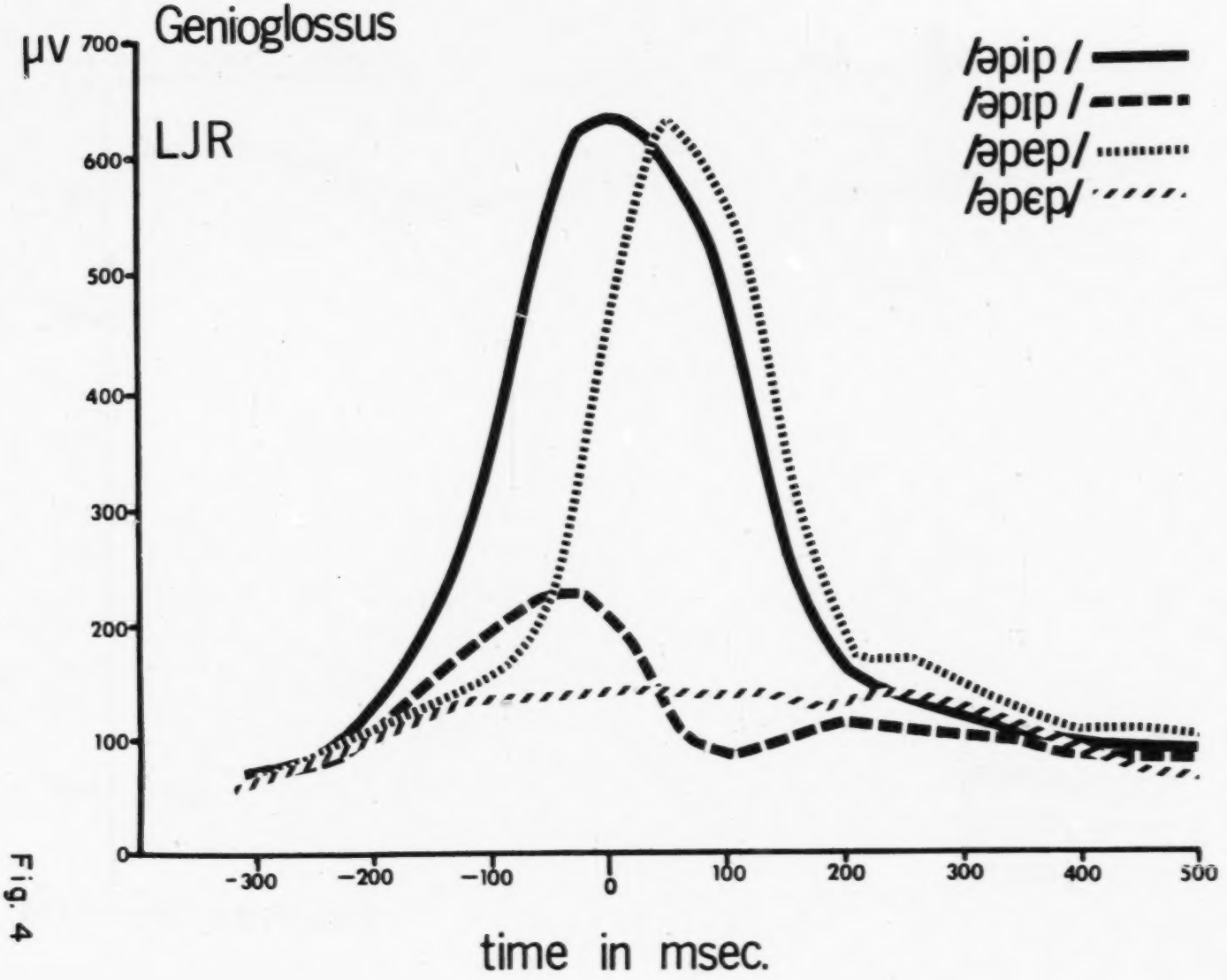

Genioglossus
μv
700
600
500
400
300
200
100
0
LJR
/əpip/
/əpɪp/
/əpep/
/əpɛp/
-300
-200
-100
0
100
200
300
400
500
time in msec.

Fig. 4

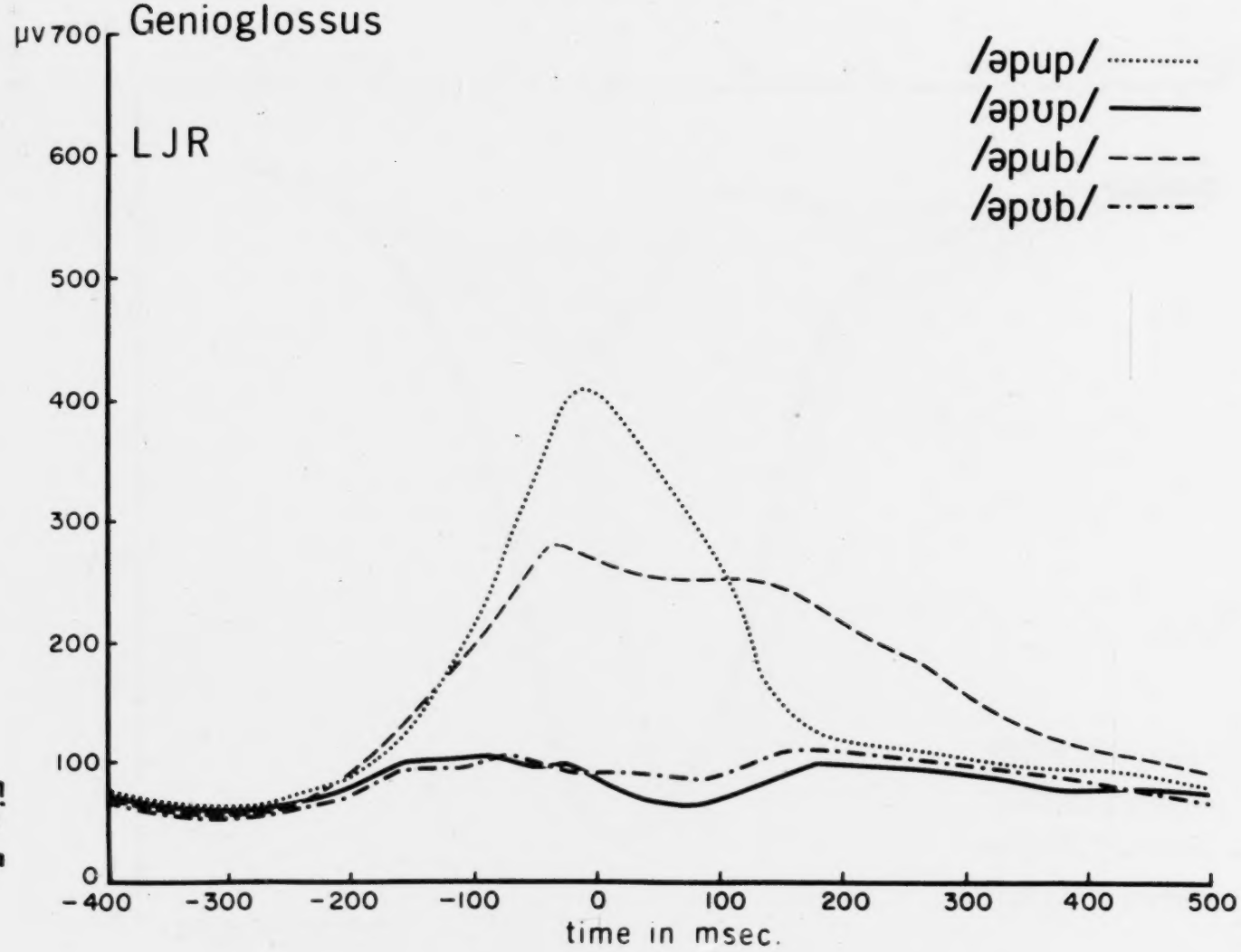

Genioglossus
μv 700
600
500
400
300
200
100
0
L J R
/əpup/
/əpʊp/
/əpub/
/əpʊb/
-400
-300
-200
-100
0
100
200
300
400
500
time in msec.

Fig. 5

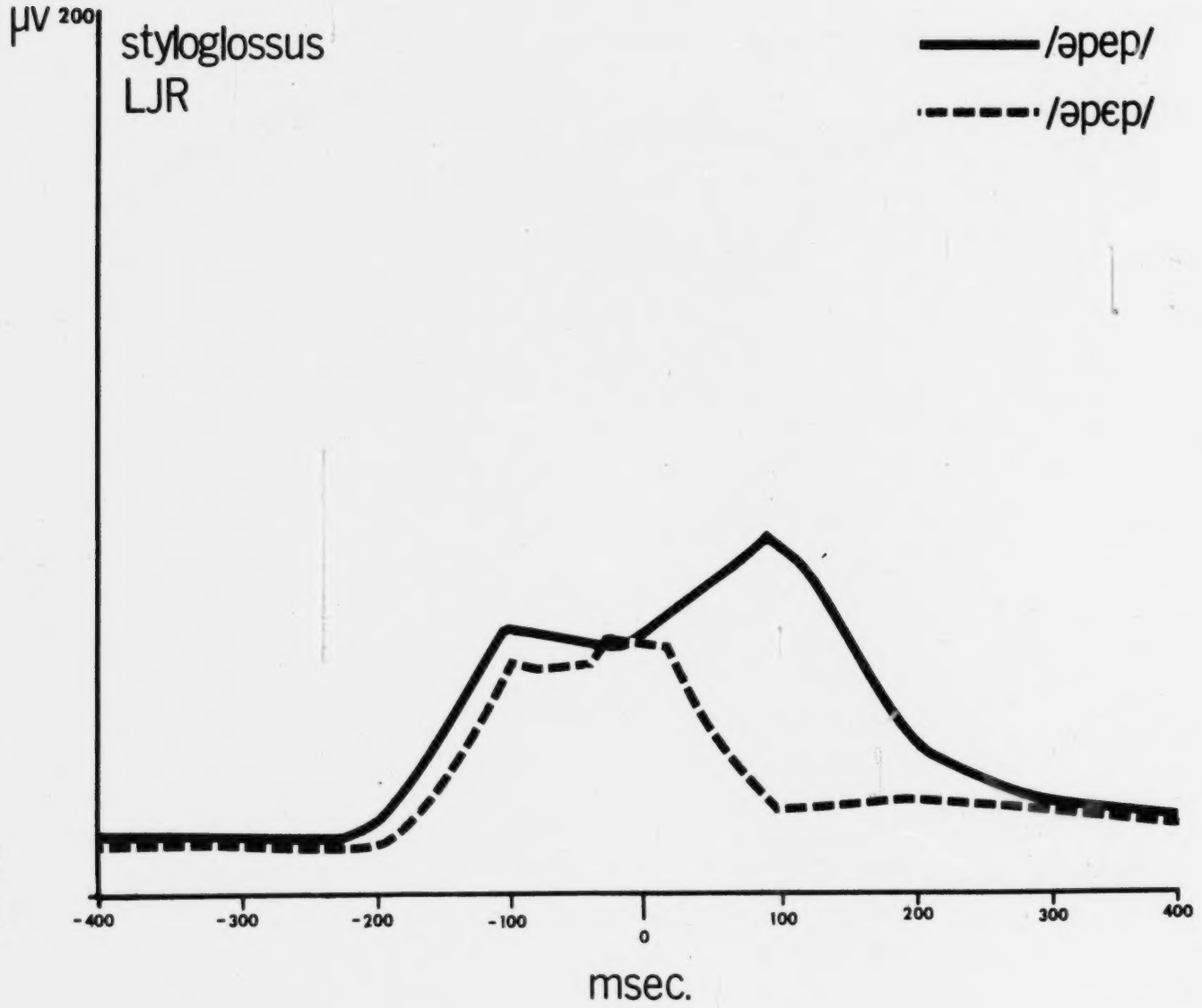

μV 200
styloglossus
LJR
/əpep/
/əpɛp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 6

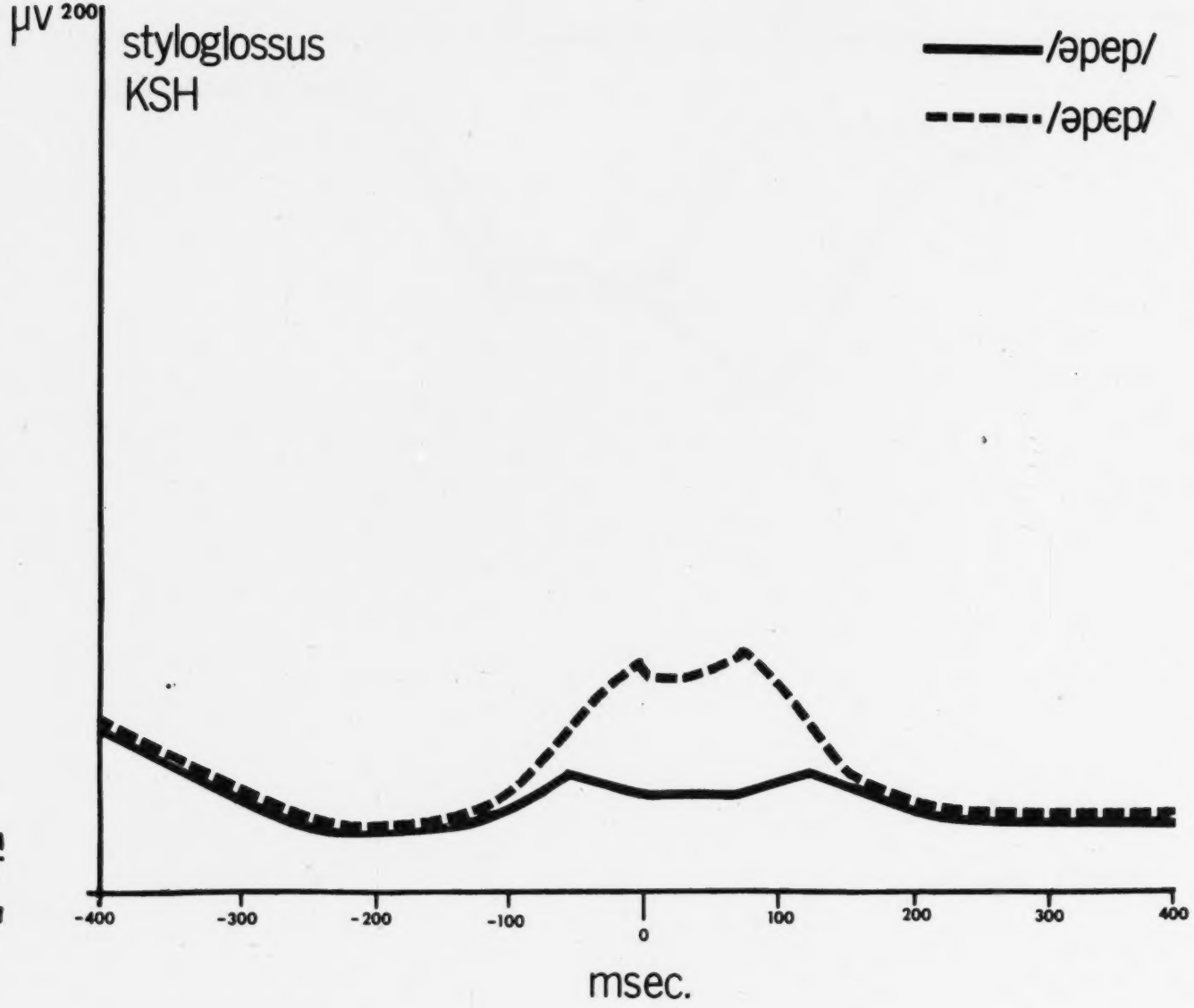

μv 200
styloglossus
KSH
/əpep/
/əpɛp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 7

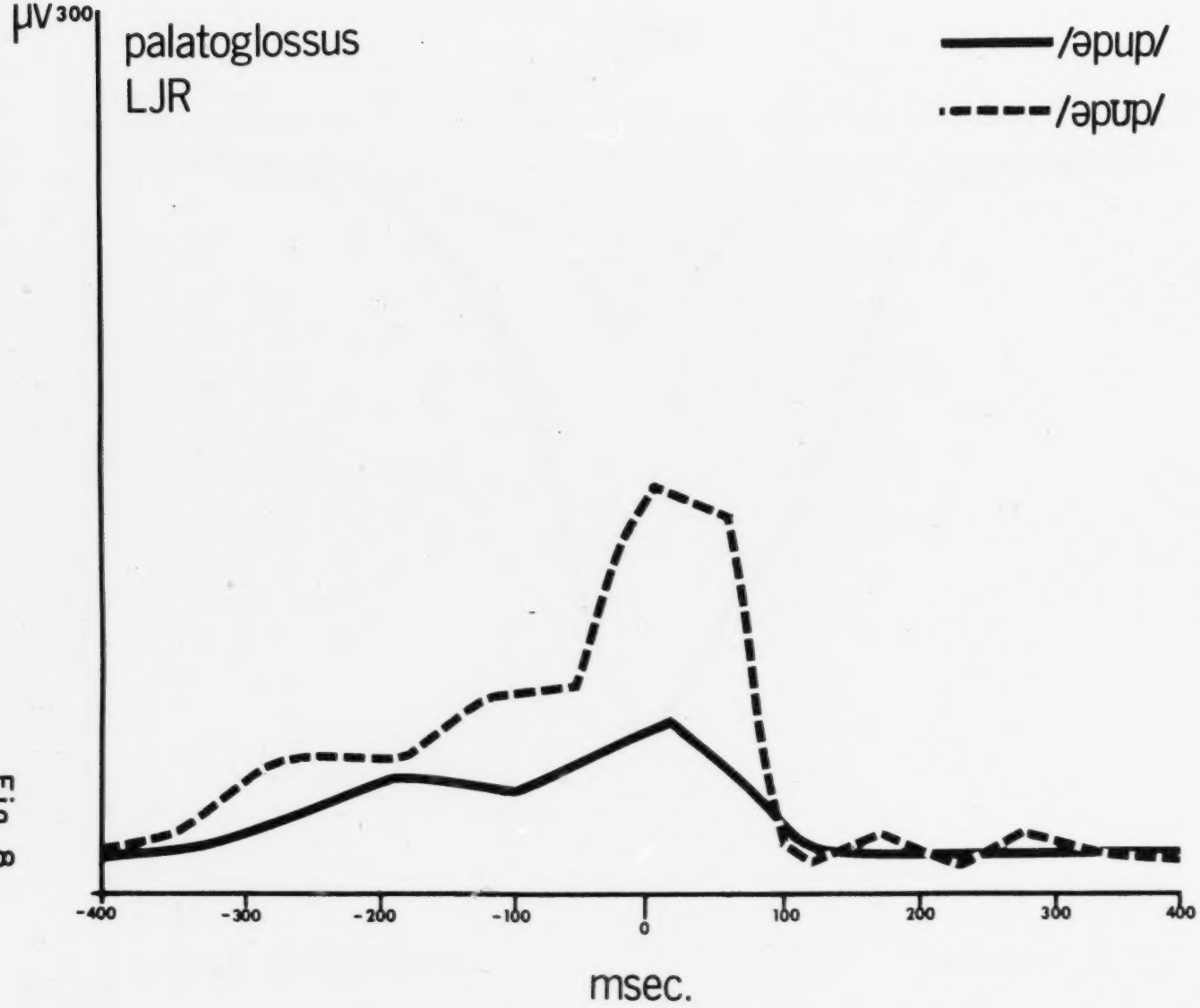

μV 300
palatoglossus
LJR
/əpup/
/əpʊp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 8

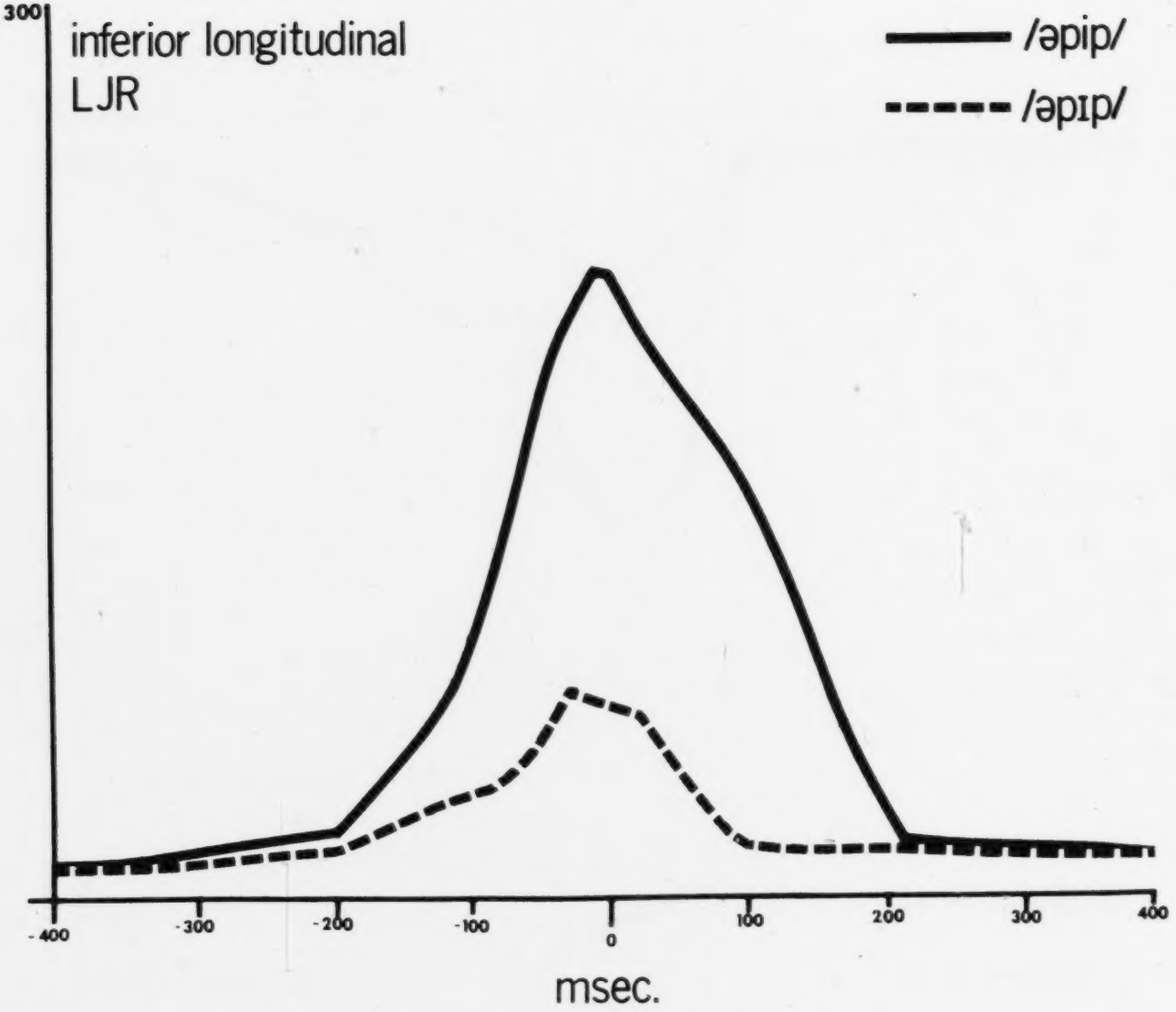

μv 300
inferior longitudinal
LJR
/əpip/
/əpɪp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 6

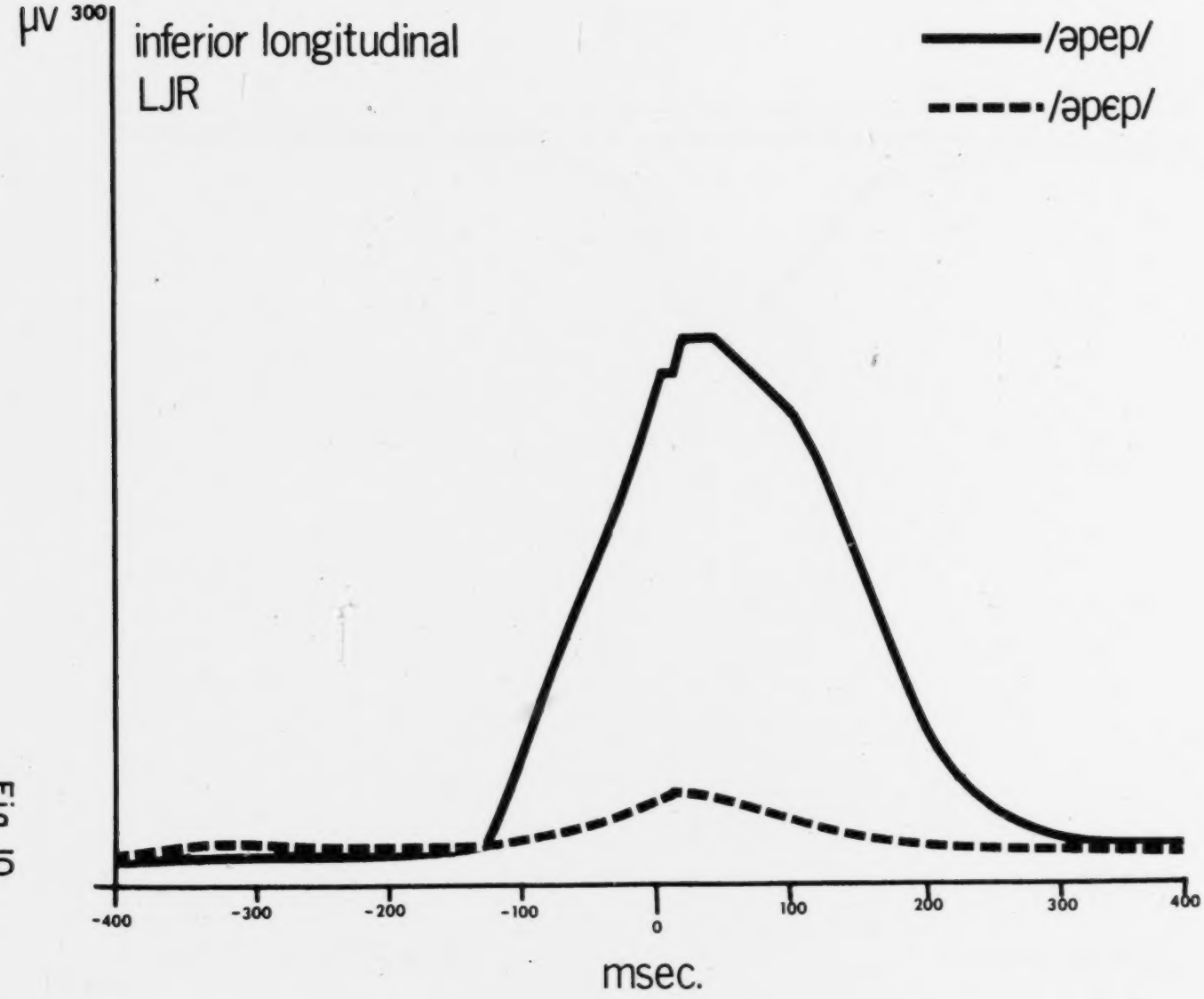

μv 300
inferior longitudinal
LJR
/əpep/
/əpɛp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 10

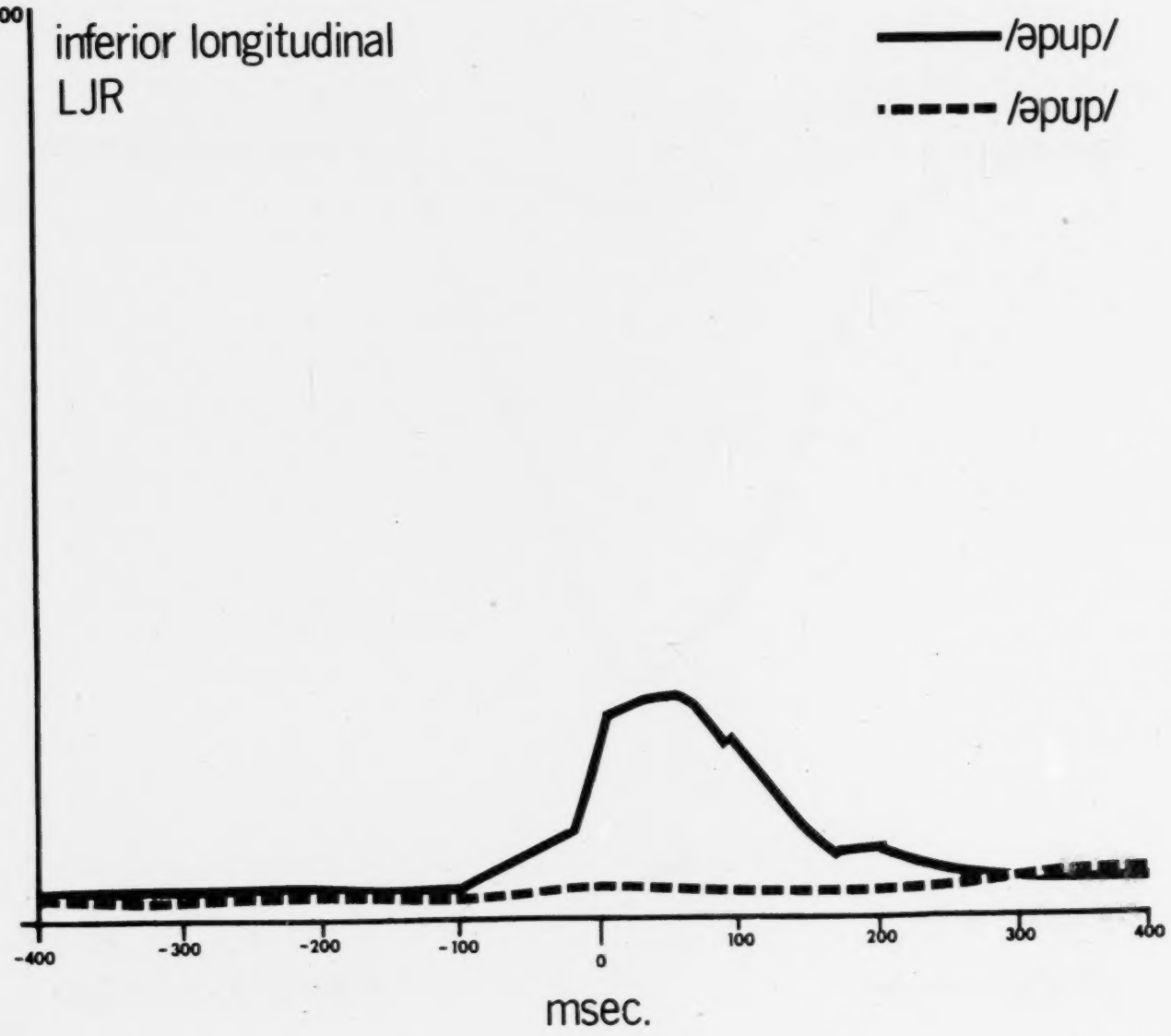

μV 300
inferior longitudinal
LJR
/əpup/
/əpup/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 11

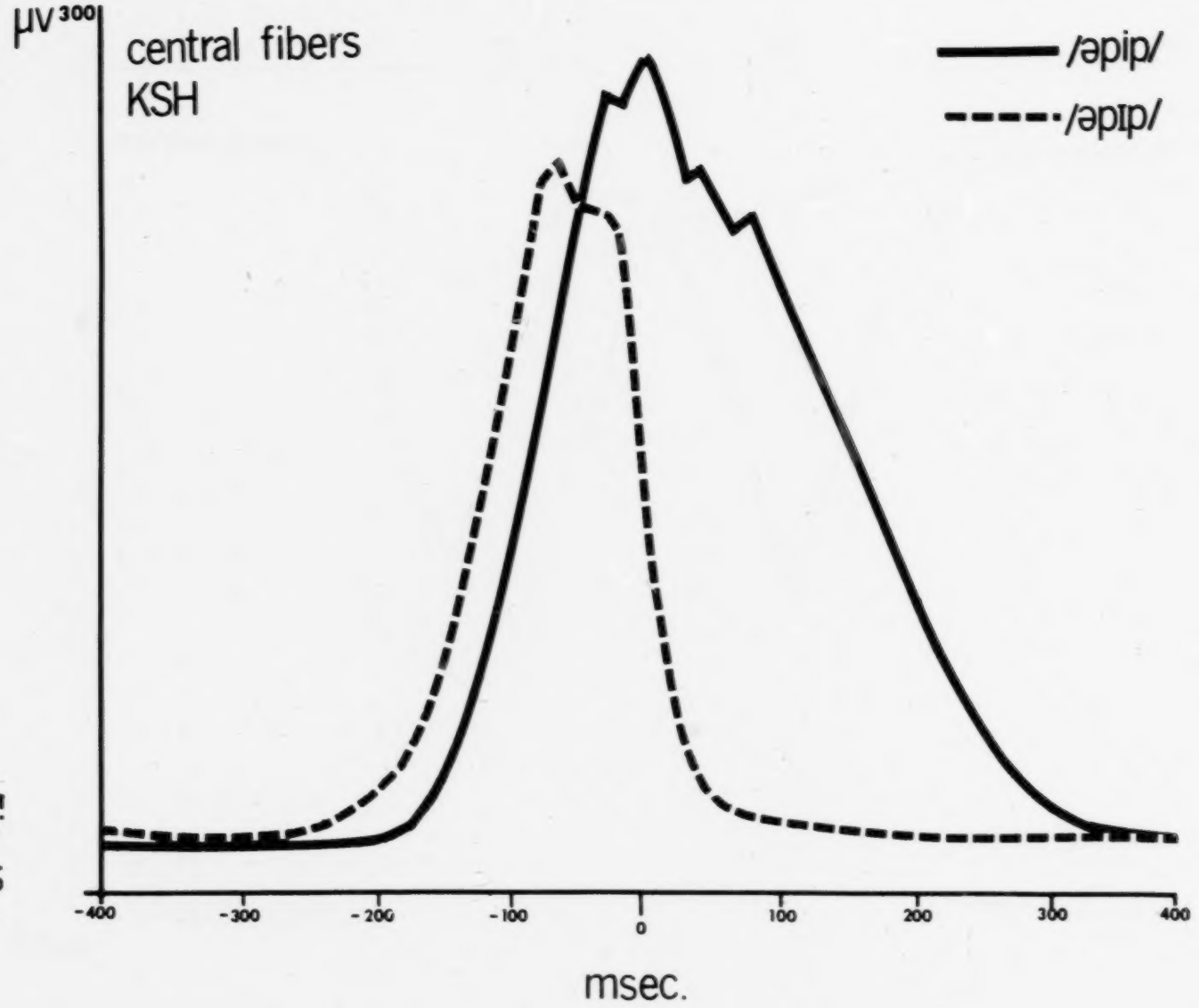

μv 300
central fibers
KSH
/əpip/
/əpɪp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 12

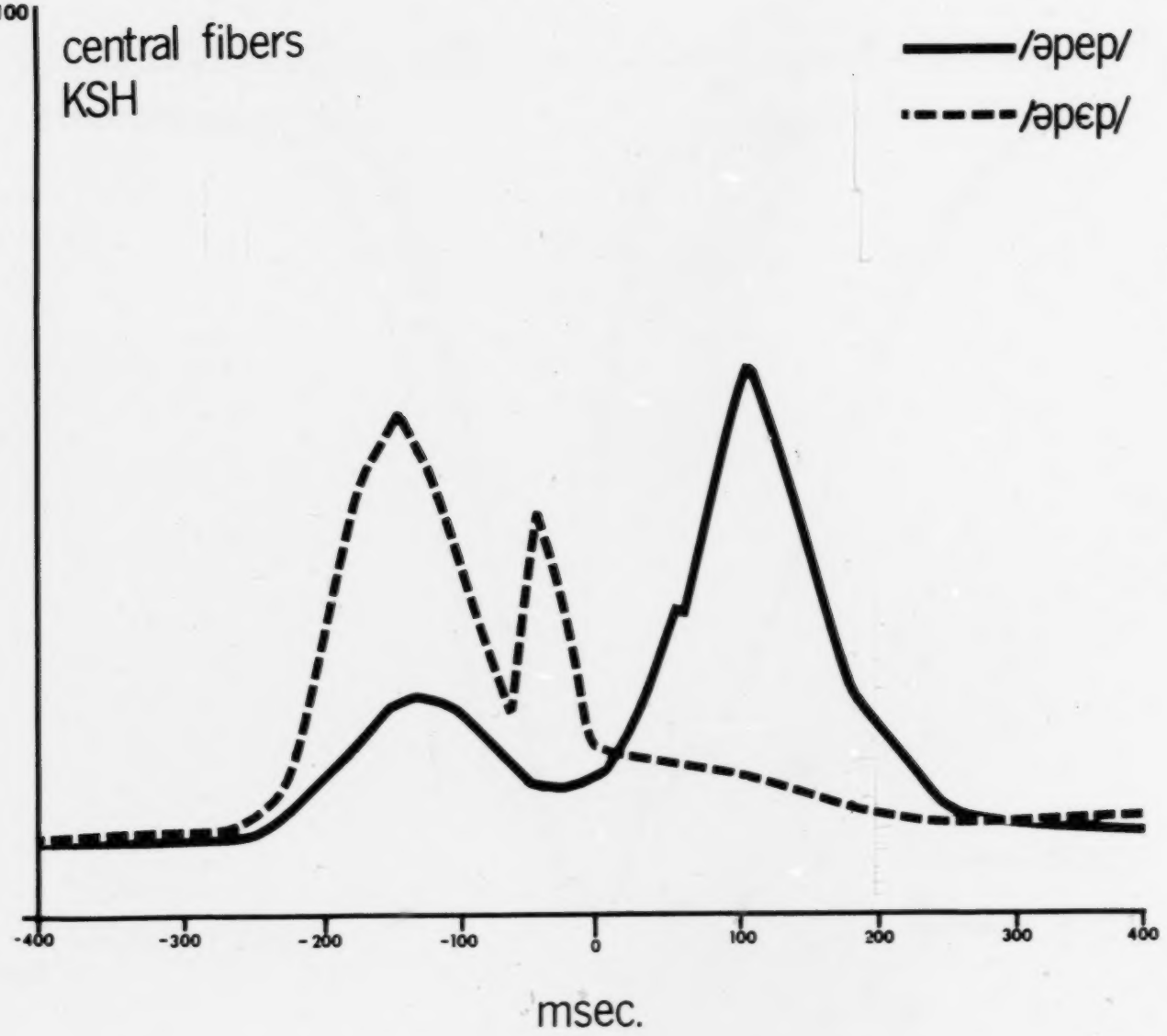

μV 100
central fibers
KSH
——/əpep/
- - - -/əpɛp/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 13

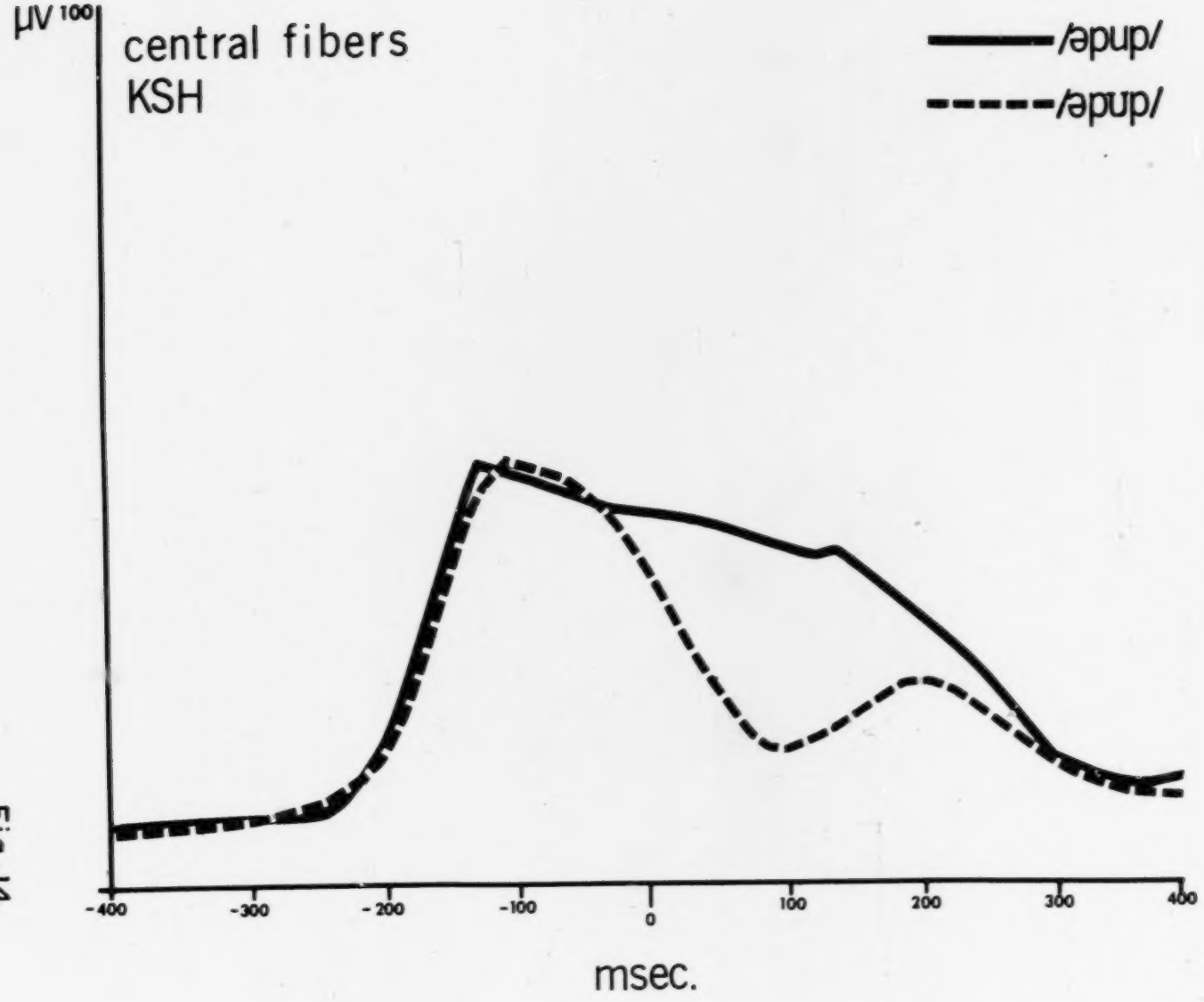

μV 100
central fibers
KSH
——/əpup/
- - - - -/əpup/
-400
-300
-200
-100
0
100
200
300
400
msec.

Fig. 14

subject number one are quite similar to each other, with easily observable, but less gross, differences between total muscular activity for tense and lax vowels than are found in subject number two. Durational differences favoring the tense vowels and bimodal curves reflecting diphthongization in the tense vowels are found in the EMG traces for the inferior longitudinal for subject number one, as they are for the genioglossus.

The data from the "central fibers" (Figures 12 and 13) are obtained only from subject number one. In general, they resemble the inferior longitudinal data. Again, the tense-lax difference follows the traditional pattern. Durational differences favoring the tense vowels are also evident, as is the diphthongization of the tense vowels. The back vowels (Figure 14) again show less muscular activity than the front, but the overall pattern is similar.

In conclusion, we find evidence for the traditionally hypothesized tense-lax difference in some muscles, namely the genioglossus, the inferior longitudinal, and the "central fibers." It must be noted, however, that when such differences occur consistently, they are always accompanied by differences in the duration of the EMG signal and often by evidence of diphthongization. For one muscle, the styloglossus, there are no consistent differences in muscular activity, and, if we are correct in our surmise concerning the inactivity of the superior longitudinal, then it is another muscle showing no consistent differences. Finally, in the palatoglossus muscle, for back vowels we found greater muscular activity for the vowel usually characterized as lax.

Thus, although we might tentatively assign the tense and lax labels to vowel categories as qualified in terms of certain muscles, the lack of consistency of the opposition in some muscles where it might be expected, and the persistence of other features which might serve as differentia, make us hesitant to claim primacy for the feature of tension in distinguishing the production of the vowel pairs investigated here.

REFERENCES

Kaplan, H. (1971) Anatomy and Physiology of Speech, 2nd ed. (New York: McGraw-Hill).

Palmer, J. (1972) Anatomy for Speech and Hearing, 2nd ed. (New York: Harper and Row).

Raphael, L. J. (1971a) An electromyographic investigation of the tense-lax feature in some English vowels. Haskins Laboratories Status Report on Speech Research SR-25/26, 131-140.

Raphael, L. J. (1971b) An electromyographic investigation of the feature of tension in some American English vowels. Haskins Laboratories Status Report on Speech Research SR-28, 179-191.

Zemlin, W. (1968) Speech and Hearing Science. (Englewood Cliffs, New Jersey: Prentice-Hall).

Effect of Speaking Rate on Labial Consonant-Vowel Articulation

T. Gay,[+] T. Ushijima,[++] H. Hirose,[+++] and F. S. Cooper
Haskins Laboratories, New Haven, Conn.

The way in which a speaker produces a given string of phones will show a good deal of variability depending upon, among other things, the suprasegmental features of stress and speaking rate. The control of speaking rate is a good illustration of the complex nature of these allophonic variations. For example, it is commonly known that during faster speech, the tongue tends to fall short of, or undershoot, vowel targets (Lindblom, 1964; Gay, 1968). This pnenomenon implies that both the rate of movement of the tongue and the activity levels of the muscles remain either unchanged, or are decreased during faster speech. However, for the production of labial consonants, another mechanism seems to operate. Here, an increase in speaking rate is accompanied by both an increase in the activity level of the muscle and an increase in the speed of movement of the articulators; these effects imply an increase, rather than a decrease, in articulatory effort (Gay and Hirose, 1973).

In an attempt to understand these phenomena further, we undertook a study of the effect of speaking rate on the coordination of lip, jaw, and tongue movements during labial consonant-vowel articulation. This experiment utilized the combined techniques of electromyography, cinefluorography, and direct view high-speed motion picture photography. This report describes some of our preliminary data.

This experiment was conducted at the Eastman Dental Center, Rochester, New York, using the cinefluorographic equipment of Dr. J. Daniel Subtelny. Hooked-wire electrodes were inserted into the orbicularis oris, anterior belly of the digastric, internal pterygoid, genioglossus, and upper longitudinal and intrinsic medial tongue muscles of two male subjects. The EMG data along with a sequence of octal code pulses were recorded on a portable tape recorder system.

Lateral view cinefluorographic films were recorded with a 16 mm cine camera set up to run at 64 fps. The X-ray generator delivered 1 msec pulses to a 9 inch image intensifier tube. A barium sulfate paste was used as an opaque medium on the tongue, and tantalum was applied along the midline of the nose, lips, and jaw

---

[+]Also University of Connecticut Health Center, Farmington.

[++]On leave from University of Tokyo.

[+++]Also University of Tokyo.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

to outline those structures. The X-ray film records were synchronized with the other records by a pulse train generated by the X-ray camera and recorded on the data tape.

High-speed motion pictures on lip movements were recorded with a 16 mm Milliken camera, set up to run at 128 fps. The motion picture and EMG data were synchronized by an annotation system that displayed octal code pulses on an LED device placed in the path of the camera. A more detailed description of the recording system can be found elsewhere (Gay and Hirose, 1973).

The speech material consisted of the consonants /p w/ and the vowels /i a u/ in a trisyllable nonsense word of the form /k V_1 C V_2 pə/, where V_1 and V_2 were all possible combinations of /i a u/ and C was either /p/ or /w/. The first two syllables were spoken with equal stress while the last was unstressed. The carrier phrase, "It's a...," preceded each utterance.

Each of the two subjects repeated each of the utterances at both normal and fast speaking rates. The EMG data were analyzed by computer (Port, 1971) and the X-ray and direct view films were analyzed by frame-by-frame measurements. The results described below are based on preliminary analysis of the data from one of the subjects.

Lip Movements

The electromyographic and direct high speed motion picture measurements for one set of /p/ utterances are illustrated in Figures 1 and 2. Figure 1 shows the averaged EMG curves for the orbicularis oris muscle for both speaking rate conditions. It is clear that the muscle activity levels for the faster speaking rate condition are substantially higher than those for the normal speaking rate condition. These differences, which occurred for both /p/ and /w/ and across all vowel environments, indicate that the completion of the lip closing gesture during faster speech requires greater articulatory effort. This increase in muscle activity is also reflected in faster rates of lip movement during faster speech (Figure 2). The faster rates of lip movement are evident for both the closing and opening phases of the /p/ gesture and often carry over to the following vowel as overshoot in lip opening. These results support our earlier data (Gay and Hirose, 1973), and again demonstrate that for the production of labial consonants during faster speech, the gesture is reorganized in terms of greater articulatory effort to enable the articulators to reach and remain at their targets.

Tongue Movement

Figures 3 and 4 illustrate the electromyographic and cinefluorographic measurements of tongue movement. Figure 3 shows the averaged curves of the genioglossus muscle for the same /VpV/ set as before. Here, as compared to the orbicularis oris, the effect of an increase in speaking rate is a substantial decrease in the activity level of the muscle. The differences in EMG activity shown here were consistent for all the other consonant-vowel combinations where the genioglossus muscle is active.

A decrease in genioglossus activity is consistent with the view that the tongue undershoots its vowel targets during faster speech. This effect is borne out by the X-ray data (Figure 4). This figure shows the position of the tongue

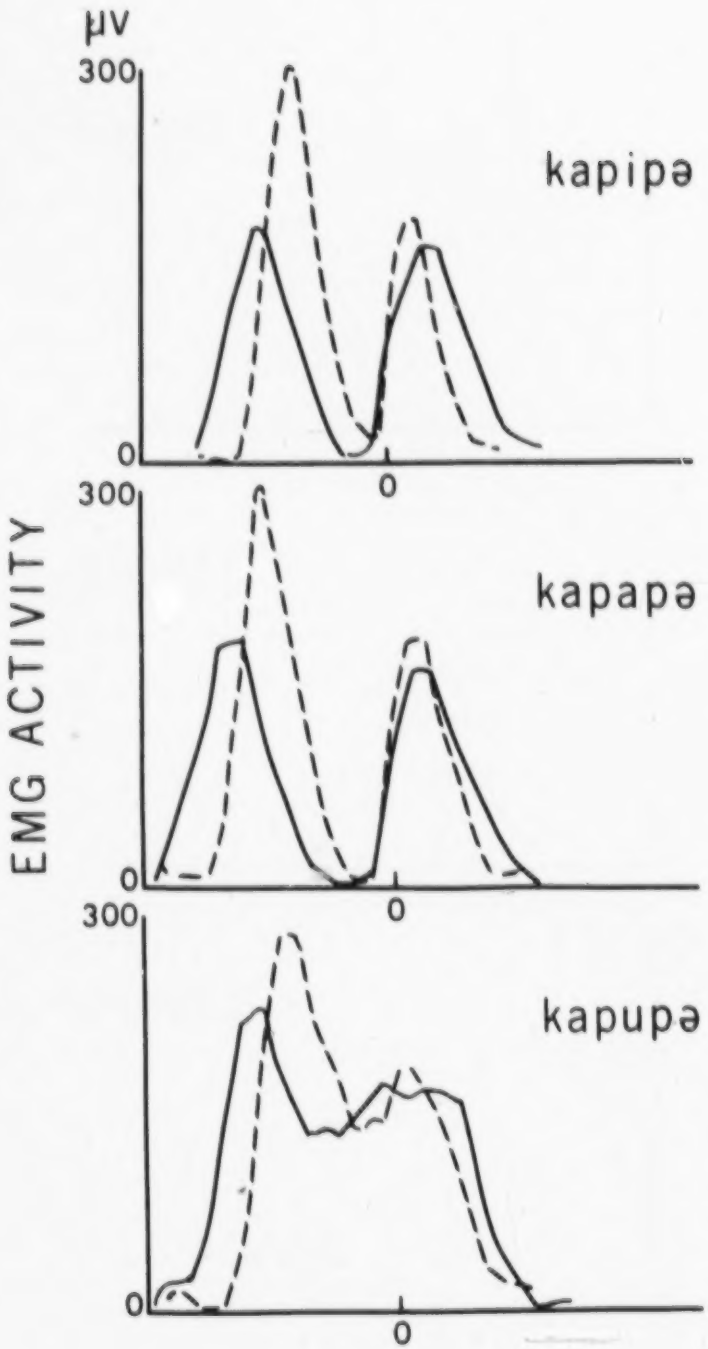


Figure 1: Averaged EMG curves for orbicularis oris muscle. "0" = offset of voicing for the second vowel. Dashed lines represent fast speaking rate, solid lines represent slow speaking rate.

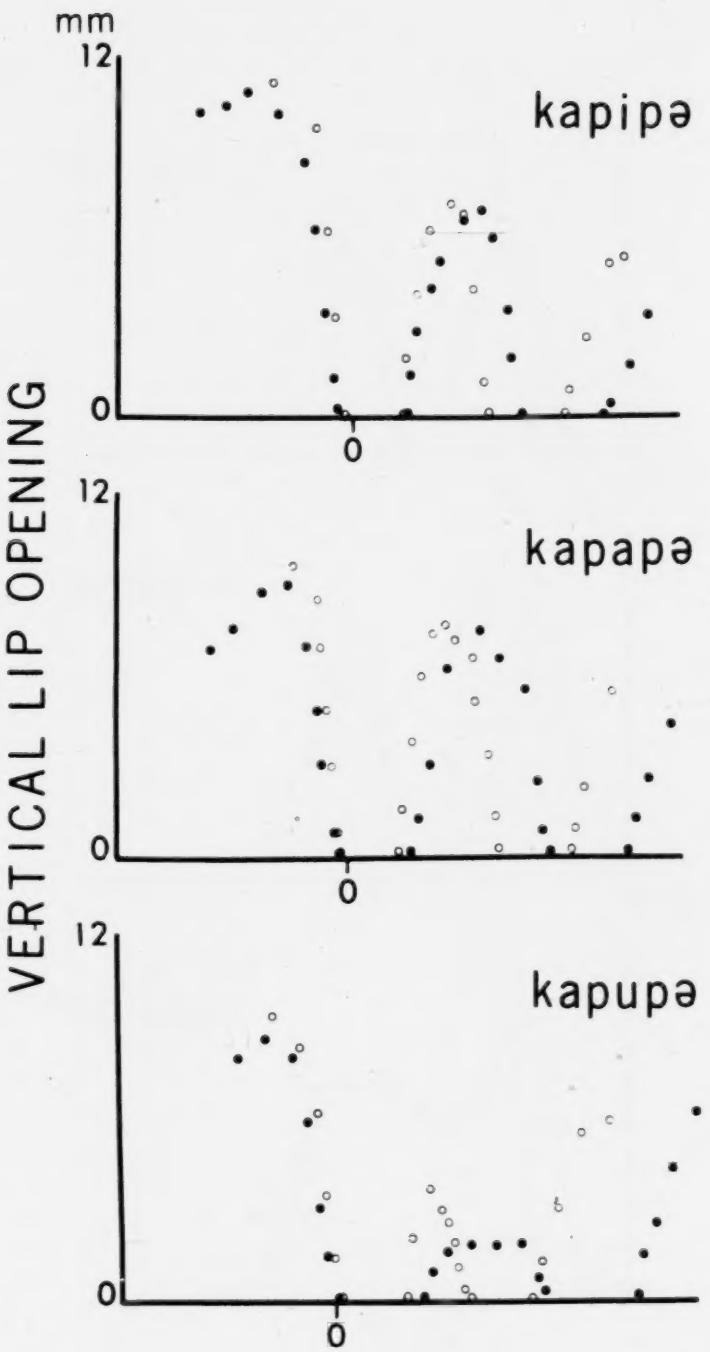


Figure 2: Vertical lip height measurements. "0" = point of /p/ closure. Filled circles represent slow speaking rate, unfilled circles represent fast speaking rate.

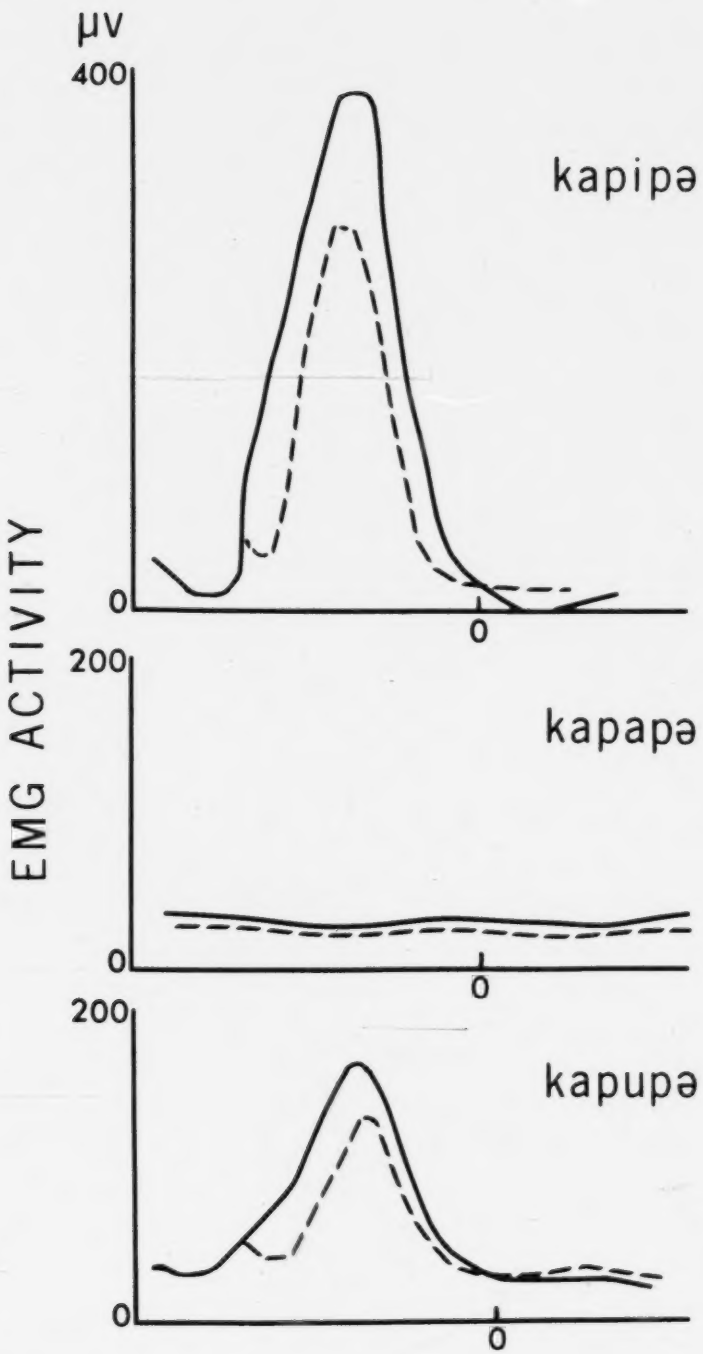


Figure 3: Averaged EMG curves for the genioglossus muscle. "0" = offset of voicing for the second vowel. Solid lines represent slow speaking rate, dashed lines represent fast speaking rate.

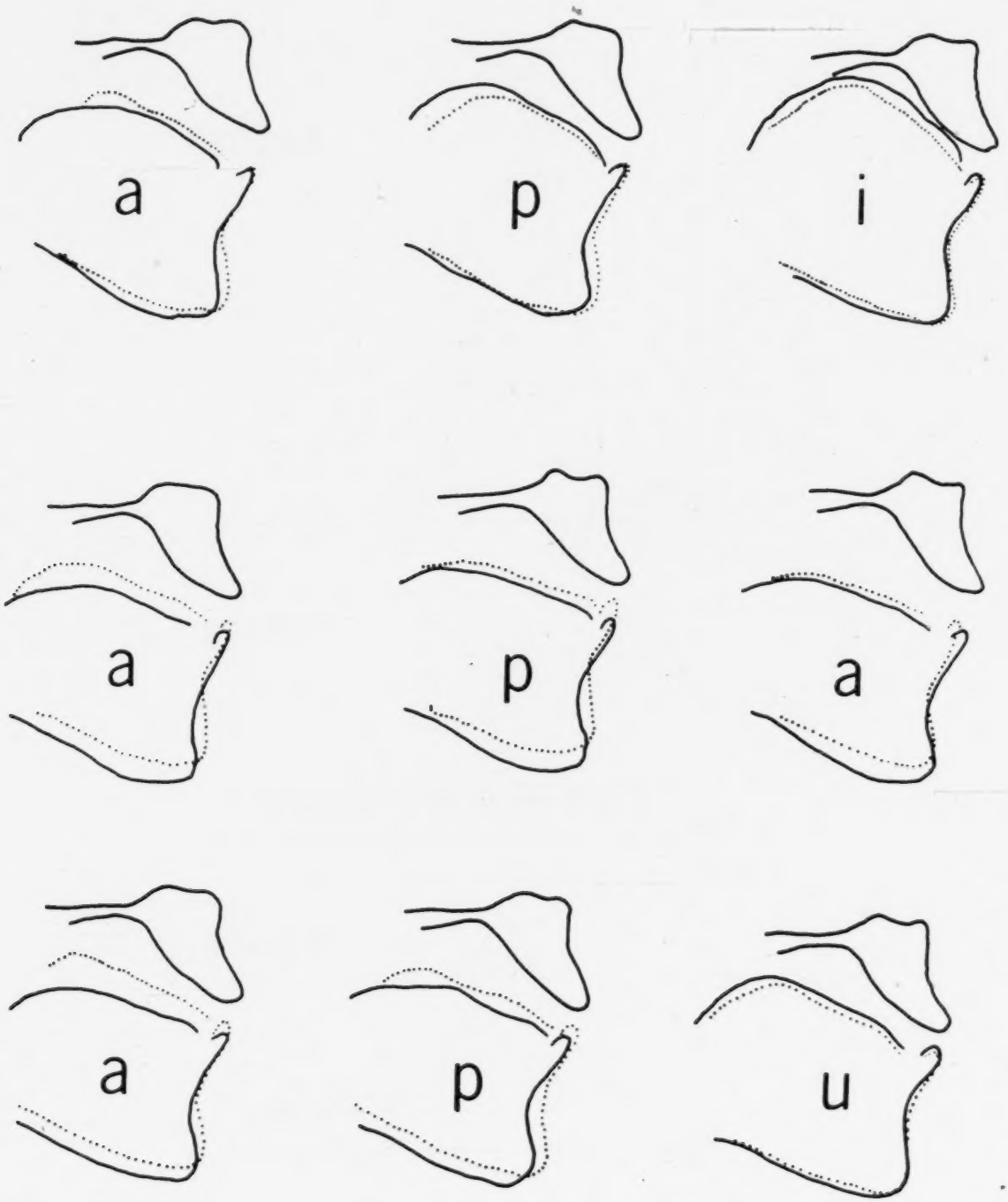


Figure 4: X-ray tracings for V_1 target, point of /p/ closure, V_2 target, for slow (solid) and fast (dotted) speaking rates.

at the V_1 target, at the time of /p/ closure, and at the V_2 target, for both speaking rate conditions. It is quite clear from this figure that the tongue does not extend as far for the vowel at the fast speaking rate as it does at the slow speaking rate; in other words, it does undershoot its target. Tongue undershoot is also consistent, occurring for all vowels in both pre- and post-consonantal positions.

The preliminary data described in this report show that the control of speaking rate cannot be accounted for by a simple quantitative model that is based on timing changes alone. This is not unexpected for stop consonant production because the lips cannot in a strict sense undershoot their targets. Rather, it is more logical that the muscles would work harder to complete the gesture under the constraint of reduced duration.

The tongue data are not so straightforward. Although the tongue does, indeed, undershoot its intended target during faster speech, it also seems that there is less muscular driving force behind the movement. Lindblom's (1963) original hypothesis, of course, is that changes in the timing of commands to the muscles, alone, are the physiological correlates of faster speech. These data indicate that there is less muscle activity, and hence, some reorganization in the opposite way, that is, some kind of programmed undershoot.

These data further suggest that the speaking rate control mechanism might be organized, at least in part, according to a muscle or muscle system rather than a phoneme system. However, additional data are required to determine if such a muscle system could be reduced, or even correlated, to phonetic or physiological features.

REFERENCES

Gay, T. (1968) Effect of speaking rate on diphthong formant movements. J. Acoust. Soc. Amer. 44, 1570-1573.

Gay, T. and H. Hirose. (1973) Effect of speaking rate on labial consonant production. Phonetica 27, 44-53.

Lindblom, B. (1963) Spectrographic study of vowel reduction. J. Acoust. Soc. Amer. 35, 1773-1781.

Lindblom, B. (1964) Articulatory activity in vowels. (Stockholm, Sweden: Speech Transmission Laboratory, Royal Institute of Technology) QPSR, STL/RIT. 2, 1-5.

Port, D. (1971) The EMG data system. Haskins Laboratories Status Report on Speech Research SR-25/26, 67-72.

On the Evolution of Language: A Unified View*

Philip Lieberman[+]
Haskins Laboratories, New Haven, Conn.

Language can be operationally defined as a communications system that permits the exchange of new, unanticipated information. Different forms of language appear to have been present in earlier stages of hominid evolution. Human language is unique, at the present time, since it makes use of "encoded" speech to achieve a rapid transfer of information. The supralaryngeal vocal tract of modern Homo sapiens is a useful factor in this encoding process which also involves special neural mechanisms. Other factors like cognitive ability and "automatization" are also necessary for language. These factors are, however, important for many aspects of human and nonhuman behavior besides language.

The evolution of language appears to have been a gradual process that first led to systems that relied on mixed gestural and vocal communication. Some hominids appear to have retained this system until comparatively recent times. Other hominids appear to have placed a greater reliance on vocal communication. Reconstructions of fossil supralaryngeal vocal tracts show that some forms, Australopithecines and "classic" Neanderthal, lacked the supralaryngeal vocal tract that is necessary for the production of fully encoded human speech. Other fossil forms, Steinheim and Es-Skhūl V, had functionally modern vocal tracts. Others, like Broken Hill, represent intermediate forms. The evolution of human language can be viewed as a three-stage process that involved (a) increased reliance on vocal communication in activities like hunting, (b) the enhancement of the vocal repertoire with the evolution of the human supralaryngeal vocal tract which produces acoustic signals that are both more distinct and more resistant to articulatory errors, and (c) the evolution of neural mechanisms that made use of the preadapted properties of the supralaryngeal vocal tract for rapid, encoded speech communication.

---

*Invited paper for presentation at the IXth International Congress of Anthropological and Ethnological Sciences, Chicago, Ill., September 1973; to be published in the Congress proceedings.

[+]Also University of Connecticut, Storrs.

[HASKINS LABORATORIES: Status Report on Speech Research SR-33 (1973)]

I shall attempt to develop a unified theory for the evolution of human language in this paper. Though this theory centrally involves the comparative, ontogenetic, and evolutionary studies of speech production with which my colleagues and I are closely identified, it also crucially involves the consideration of other recent, and not so recent, studies of cognitive ability in nonhuman primates: hunting, bipedal posture, the neural correlates of auditory perception, visual perception in adult and infant humans, speech perception in humans, play activity, gesture, etc. In short, I shall attempt to synthesize a great deal of data into what I hope is a coherent and testable theory. Like all theories it cannot account for everything. This theory does, however, appear to explain and relate a number of phenomena that otherwise seem quite unrelated. Moreover, it appears to point to a coherent evolutionary process that relates the communications systems of other animals to human language. Most importantly, it points out a number of questions that can be resolved through controlled experiments and careful observations.

I have drawn on a number of seemingly disparate ethological, anatomical, psychological, and anthropological sources because I think that it is obvious that no single factor is, in itself, responsible for the evolution of human language. Evolution is a complex process that inherently involves all aspects of the life cycle and environment of a species and its relationships to other species. Though particular factors like, for example, gestural communication (Hewes, 1971) undoubtedly had an important role in the evolution of human language, no single factor can, in itself, provide the "central key" to the puzzle. Everything depends on everything else and the interaction is the crucial factor if anything is. Gestural communication, for example, depends on the prior existence of visual pattern recognition, analysis, cognitive ability, and bipedal posture. Visual pattern identification probably depends, in turn, on natural selection for visual ability in an arboral environment. Bipedal posture, in turn, probably again depends on prior selection for brachiation in an arboral environment (Campbell, 1966).

Note that I am not saying that we cannot analyze the factors that underlie the evolution of human language. I am proposing that the process involved many factors. One of these factors appears to be the process of "preadaptation," that is, natural selection channeled development in particular directions because of previous modifications selected for some other role. Darwin's (1859) comments concerning the evolution of the lung from the swim bladder describe one of the first and most convincing examples of preadaptation.

Let me begin by listing the evolutionary factors that I will discuss in this paper. There probably are more factors but I propose that these are the central factors in the evolution of human language. I shall order the factors in terms of their probable role in differentiating the language of modern man from progressively earlier hominids and other animals. In other words, I shall first list the factors that I think were most important in the late stages of human evolution and proceed to factors that probably were more important in earlier stages. It is important to note that I am not categorically differentiating human language, i.e., the language of present-day Homo sapiens, from other languages, e.g., the possible language of present-day chimpanzees.

Linguists have been somewhat anthrocentric in defining language to be necessarily human language. I will define a language to be a communications system

that is capable of transmitting new information. In other words, I am operationally defining language as a communications system that places no inherent restriction on the nature or quality of the information transferred. Clearly this definition does not require that all languages have all of the properties of human language.

Factor I - Specialized Speech Encoding and Decoding

Modern man's communications achieve a high rate of transmission speed by means of a process of speech encoding and decoding. The rate at which meaningful sound distinctions are transmitted in human speech is about twenty to thirty segments per second. That is, phonetic distinctions that differentiate meaningful words, e.g., the sounds symbolized by the symbols [b], [æ], and [t] in the word bat, are transmitted, identified, and sorted at a rate of twenty to thirty segments per second. It is obvious that human listeners cannot simply transmit and identify these sound distinctions as separate entities. The fastest rate at which sounds can be identified is about seven to nine segments per second. Sounds transmitted at a rate of twenty per second indeed merge into an undifferentiable tone. The linguist's traditional conception of phonetic elements comprising a set of "beads on a string" clearly is not correct at the acoustic level. How, then, is speech transmitted and perceived?

The results of the past twenty years of research on the perception of speech by humans demonstrated that the individual sounds like [b], [æ], and [t] are encoded, that is, "squashed together," into the syllable-sized unit [b æ t] (Liberman, Cooper, Shankweiler, and Studdert-Kennedy, 1967). A human speaker in producing this syllable starts with his supralaryngeal vocal tract, i.e., his tongue, lips, and velum, etc., in the positions characteristic of [b]. He, however, does not maintain this articulatory configuration but instead moves his articulators towards the positions that would be attained if he were instructed to maintain an isolated, steady [æ]. He never reaches these positions, however, because he starts towards the articulatory configuration characteristic of [t] before he ever reaches the "steady state" (isolated and sustained) vowel [æ]. The articulatory gestures that would be characteristic of each isolated sound are never attained. Instead, the articulatory gestures are melded together into a composite, characteristic of the syllable.

The sound pattern that results from this encoding process is itself an indivisible composite. Just as there is no way of separating with absolute certainty the [b] articulatory gestures from the [æ] gestures (you can't tell exactly when the [b] ends and the [æ] begins), there is no way of separating the acoustic cues that are generated by these articulatory maneuvers. The isolated sounds have a psychological status as motor control or "programming" instructions for the speech production apparatus. The sound pattern that results is a composite and the acoustic cues for the initial and final consonants are largely transmitted as modulations imposed on the vowel. The process is, in effect, a time compressing system. The acoustic cues that characterize the initial and final consonants are transmitted in the time slot that would have been necessary to transmit a single isolated [æ] vowel.

The human brain decodes, that is, unscrambles the acoustic signal in terms of the articulatory maneuvers that were put together to generate the syllable. The individual consonants [b] and [t], though they have no independent acoustic

status, are perceived as discrete entities. The process of human speech perception inherently requires "knowledge" of the acoustic consequences of the possible range of human supralaryngeal vocal tract speech articulation (Liberman et al., 1967; Lieberman, 1970, 1972). The special speech processing involved appears to involve crucially the dominant hemisphere of the human brain (Kimura, 1964; Liberman et al., 1967; Shankweiler and Studdert-Kennedy, 1967). We will discuss the process of human speech perception in more detail with respect to its interrelation with the anatomy of the human vocal tract. For the moment, we will note that the special neural devices necessary for the decoding of human speech may be comparatively recent evolutionary acquisitions.

Factor 2 - Special Supralaryngeal Vocal Tract Anatomy

Modern man's speech-producing apparatus is quite different from the comparable systems of living nonhuman primates (Lieberman, 1968; Lieberman, Klatt, and Wilson, 1969; Lieberman, Crelin, and Klatt, 1972a). Nonhuman primates have supralaryngeal vocal tracts in which the larynx exits directly into the oral cavity (Negus, 1949). In the adult human the larynx exits into the pharynx. The only function for which the adult human supralaryngeal vocal tract appears to be better adapted is speech production. Understanding the anatomical basis of human speech requires that we briefly review the source-filter theory of speech production (Chiba and Kajiyama, 1958; Fant, 1960). Human speech is the result of a source, or sources, of acoustic energy being filtered by the supralaryngeal vocal tract. For voiced sounds, that is, sounds like the English vowels, the source of energy is the periodic sequence of puffs of air that pass through the larynx as the vocal cords (folds) rapidly open and shut. The rate at which the vocal cords open and close determines the fundamental frequency of phonation. Acoustic energy is present at the fundamental frequency and at higher harmonics. The fundamental frequency of phonation can vary from about 80 Hz for adult males to about 500 Hz for children and some adult females. Significant acoustic energy is present in the harmonics of fundamental frequency to at least 3000 Hz. The fundamental frequency of phonation is, within wide limits, under the control of the speaker who can produce controlled variations by varying either pulmonary air pressure or the tension of the laryngeal muscles (Lieberman, 1967). Linguistically significant information can be transmitted by means of these variations in fundamental frequency as, for example, in Chinese where these variations are used to differentiate different words.

The main source of phonetic differentiation in human languages, however, arises from the dynamic properties of the supralaryngeal vocal tract acting as an acoustic filter. The length and shape of the supralaryngeal vocal tract determines the frequencies at which maximum energy will be transmitted from the laryngeal source to the air adjacent to the speaker's lips. These frequencies at which maximum acoustic energy will be transmitted are known as formant frequencies. A speaker can vary the formant frequencies by changing the length and shape of his supralaryngeal vocal tract. He can, for example, drastically alter the shape of the airway formed by the posterior margin of his tongue body in his pharynx. He can raise or lower the upper boundary of his tongue in his oral cavity. He can raise or lower his larynx and retract or extend his lips. He can open or close his nasal cavity to the rest of the supralaryngeal vocal tract by lowering or raising his velum. The speaker can, in short, continually vary the formant frequencies generated by his supralaryngeal vocal tract. The acoustic properties that, for example, differentiate the vowels [a] and [i] are determined solely by the shape and length differences that the speaker's supralaryngeal

vocal tract assumes in articulating these vowels. The situation is analogous to the musical properties of a pipe organ where the length and type (open or closed end) of pipe determines the musical quality of each note. The damped resonances of the human supralaryngeal vocal tract are, in effect, the formant frequencies. The length and shape (more precisely the cross-sectional area as a function of distance from the laryngeal source) determine the formant frequencies.

The situation is similar for unvoiced sounds where the vocal cords do not open and close at a rapid rate, releasing quasiperiodic puffs of air. The source of acoustic energy in these instances is the turbulence generated by air rushing through a constriction in the vocal tract. The vocal tract still acts as an acoustic filter but the acoustic source may not be at the level of the larynx as, for example, in the sound [s] where the source is the turbulence generated near the speaker's teeth.

The anatomy of the adult human supralaryngeal vocal tract permits modern man to generate supralaryngeal vocal tract configurations that involve abrupt discontinuities at its midpoint. These particular vocal tract shapes produce vowels with unique acoustic properties like [a], [i], and [u], as well as consonants like [g] and [k]. The acoustic properties of these particular sounds will be discussed in detail, but for the moment I will simply note that these sounds minimize the problems of precise articulatory control. A speaker can produce about the same formant frequencies for an [i], for example, while he varies the position of the midpoint area function discontinuity by one or two centimeters (Stevens, in press). They are also sounds that are maximally distinct acoustically. Moreover, they are sounds a human listener can efficiently use to establish the size of the supralaryngeal vocal tract that he is listening to. This last property relates to Factor 1, the specialized speech encoding and decoding that characterizes human language. The reconstructions of the supralaryngeal vocal tracts of various fossil hominids that my colleague Edmund S. Crelin has made indicate that some extinct hominids lacked the anatomical basis for producing these sounds, while other hominids appear to have the requisite anatomical specializations for human speech. I will, of course, return to this topic.

Factor 3 - Automatization and Cognitive Ability

There are two interrelated aspects to the cognitive abilities that underlie language. One is the process that I will term automatization. Human language involves rapidly executing complex sequences of articulatory maneuvers or making equally complex perceptual decisions regarding the identity of particular sound segments. At a higher level, complex phonologic and morphophonemic relationships must be determined. None of these processes is, however, what the speaker or listener is directly concerned with. The semantic content of the message is the primary concern of the speaker or listener. The sending and receiving processes are essentially automatic. No conscious thought is expended in the process of speech production, speech perception, or any of the syntactic or morphophonemic stages that may intervene between the semantic content of the message and the acoustic signal. It is clear that "automatized" skills are not unique to human language. Other aspects of human activity, for example, dance, involve similar phenomena. The novice dancer must learn the particular steps and movements that characterize a particular dance form. Once the steps have been learned they must become automatized. The dance itself involves the complex sequences. Playing a musical instrument, skiing, or driving a car all involve automatized behavior.

The bases for the automatized behavior that is a necessary condition for human language may reside in cross-modal transfers from other systems of hominid and hominoid activity. Tool use, for example, requires a high degree of automatization if it is to be effective. You cannot stop to think how to use a hammer every time you drive a nail in. Hunting is perhaps a still stronger case. A successful hunter must be able to thrust his spear accurately without pausing to think about the mechanics of spear thrusting. Natural selection would quickly favor the retention of superior automatization. Automatized behavior pervades all aspects of culture. Indeed a cultural response is, to a degree, a special case of automatized behavior. In simpler animals cultural responses are perhaps less subject to environmental pressures. In humans they may be more subject to external forces rather than innate mechanisms, but they are no less automatized once learned.

A special factor that may be germane to automatized behavior is that a "plastic" period appears to be involved. It is comparatively easy to shape behavior during the "plastic" period. Afterwards is is either impossible or relatively difficult to modify automatized behavior. Puppies thus can be trained more readily than adult dogs. All humans can readily learn different languages in their youth. Most humans can learn a foreign, i.e., unfamiliar language, only with great difficulty (or not at all) during adult life. The same comments probably apply to learning to play the violin, tight-rope walking, etc., though no definitive studies have yet been made.

Cognitive ability. Cognitive ability is a necessary factor in human language. Linguists often tend to assume that cognitive ability is linguistic ability. Indeed, since the time of Descartes the absence of human language in other animals has been cited as a "proof" of man's special status and of the lack of cognitive ability in all other species. Human language has been assumed to be a necessary condition for human thought. The absence of human language has been, conversely, assumed to be evidence of the lack of all cognitive ability.

It is clear that cognitive, i.e., logical, abilities can be demonstrated or observed in many animals. Behavioral conditioning, for example, which can be applied with great success to pigeons and rats, itself can be viewed as a demonstration of logical ability on the part of the "conditioned" animal. Pavlov's dogs had to make a logical association between the bell and food. Calling the animal's response a conditioned reflex obscures the fact that the animal had to be able logically to connect the sound of the bell with food. The same "conditioned" response often can be observed as a human gourmet regards the menu. In both cases cognitive ability must interpose between the token of the food that is anticipated and the observed physiologic response. The human gourmet is hopefully more flexible, adaptive, and discriminating than Pavlov's dogs; however, the basic process is similar. In Homo sapiens the cognitive abilities that underlie this particular aspect of behavior are simply more complex than those of Canus familiarus.

The particular cognitive abilities that are associated with presumably "unique" human behavioral patterns like tool use have been observed in chimpanzees (Goodall, 1971) and sea otter (Kenyon, 1969). Some of the cognitive abilities that have been traditionally associated with human language have likewise been demonstrated in experiments by Gardner and Gardner (1969) and by Premack (1972). Premack's experiments, in particular, clearly demonstrate that cognitive ability and human language cannot be regarded as the same biologic ability.

Chimpanzees do not possess the phonetic apparatus of human language. They have available a subset of the phonetic distinctions that are available to modern man. Chimpanzees could, using the phonetic distinctions that are available to them, establish a language. This language's phonetic system might not be as efficient as modern man's but it could form the basis of a language (using our operational definition of language as a communications system capable of transmitting unanticipated, new knowledge). The difference, at the phonetic level, between human language and this hypothetical chimpanzee language would be quantitative rather than qualitative. Premack's experiments demonstrate that the cognitive abilities of chimpanzees are, at worst, restricted to some subset of the cognitive abilities available to humans. The difference at the cognitive level, is thus also probably quantitative rather than qualitative.

It is important to note, at this point, that quantitative functional abilities can be the bases of behavioral patterns that are qualitatively different. I think that this fact is sometimes not appreciated in discussions of gradual versus abrupt change. A modern electronic desk calculator and a large general purpose digital computer, for example, may be constructed using similar electronic logical devices and similar magnetic memories. The large general purpose machine will, however, have 1,000 to 10,000,000 times as many logical and memory devices. The structural differences between the desk calculator and general purpose machine may thus simply be quantitative rather than qualitative. The "behavioral" consequence of this quantitative difference can, however, be qualitative. The types of problems that one can solve on the general purpose machine will differ in kind, as well as in size, from those suited to the desk calculator. The inherent cognitive abilities of humans and chimpanzees thus could be quantitative and still have qualitative behavioral consequences.

The cognitive abilities that are typically associated with human language may have their immediate origins in the complex patterns of hominid behavior associated with tool use, tool making, and hunting. Hewes (1971) makes a convincing case for the role of gestural communication in the earliest forms of hominid language and associates language with the transference of cognitive ability from these complex behavioral patterns. I would agree with Hewes, but I would not limit the earliest hominid languages to gestures, nor would I restrict the cognitive abilities that underlie language to hominids. Tool use and hunting certainly are not exclusively hominid patterns of behavior.

We can get some insights on the neural abilities of nonhuman primates by noting the phylogenetic evolution of the peripheral systems involved in information gathering and communication. The acute color vision of primates, for example, would have had no selective advantage if it were not coupled with matching cognitive processes. Gestural communication is consistent with the evolution and retention of increasingly complex facial musculature in the phylogenetic order of primates. It is likewise unlikely that gestural communication was at any stage of hominid evolution the sole "phonetic" medium. Negus (1949), by the methods of comparative anatomy, demonstrates that the larynges of nonhuman primates are adapted for phonation at the expense of respiratory efficiency. The far simpler larynx of the lung fish is better adapted for respiration and protecting the lungs. Clearly mutations that decreased respiratory efficiency would not have been retained over a phylogenetic order unless they had some selectional advantage. The cognitive skills that underlie linguistic ability in hominids thus probably evolved from cognitive facilities that have a functional role in the social

behavior and communications of other animals. Like automatization, these skills would appear to be part of the biologic endowment of many species and their continued development in higher species is concomitant with behavioral complexity. The transference of these cognitive skills to human language thus could be viewed as yet another instance of "preadaptation," the use of cognitive processes for language that originally evolved because of the selective advantages conferred on activities like hunting, evading natural enemies, food gathering, etc.

The Speech Abilities of Neanderthal and other Fossil Hominids

As I noted before, it is apparent that no single factor can be in any reasonable way identified as the "key" to language. The two factors that appear to be most recent in shaping the particular form of human language are, however, Speech Encoding and Speech Producing Anatomy. Certain neural mechanisms must be present for the perception of speech (Lenneburg, 1967). It is difficult to make any substantive inferences about the presence or absence of particular neural mechanisms in the brains of extinct fossil hominids since we can deduce only the external size and shape of the brain from a fossil skull. Also, we lack a detailed knowledge of how the human brain functions. We could not really assess the linguistic abilities of a modern man simply by examining his brain. Fortunately, we can derive some insights on the nature of speech perception in various fossil hominids by studying their speech producing anatomy. The relationship between speech anatomy and speech perception is very much like the relationship between bipedalism and the detailed anatomy of the pelvic region. The anatomy is a necessary condition, though neural ability is also necessary.

The methodology that has enabled us, and I must emphasize that this research has been a joint enterprise, to reconstruct the speech producing anatomy of extinct hominids is that proposed by Charles Darwin. Darwin in Chapters X and XIII of On the Origin of Species (1859) discussed both the "affinities of extinct species to each other, and to living forms," and "embryology." We have applied the methods of comparative and functional anatomy to the speech producing anatomy of present-day apes and monkeys and to normal human newborns. We first assessed the speech producing abilities of these living animals in terms of their speech producing anatomy. We found that their supralaryngeal vocal tracts inherently restricted their speech producing abilities. We then noted that certain functional aspects of the morphology of the skulls of these living animals resembled similar features of extinct fossil hominids.

The reconstructions of the supralaryngeal vocal tracts of the La Chapelle-aux-Saints, Es-Skhūl V, Broken Hill, Steinheim, and Sterkfontein 5 fossils were made by my colleague Edmund S. Crelin by means of the homologues that exist between these skulls and living forms, the marks of the muscles on the fossil skulls, and the general methods of comparative anatomy. Crelin's (1969) previous experience with the anatomy of the newborn was especially relevant since we can see in the human newborn many of the relevant skeletal features associated with the soft tissue structures that must have occurred in certain of these now extinct hominid forms. In most cases we made use of casts of the fossil material made available by the Wenner-Gren Foundation. For the La Chapelle-aux-Saints and Steinheim fossils, casts made available by the University Museum, Philadelphia, Pennsylvania, were employed. The original La Chapelle-aux-Saints fossil as well as the La Ferrasie and La Quina child's fossil were also examined with the cooperation of the Musée de l'Homme in Paris and the Musée des Antiquites

Nationales in St. Germain-en-Laye. We attempted to examine the original Steinheim fossil but were not successful. The details of the reconstructions are discussed in our published and forthcoming papers (Lieberman and Crelin, 1971; Lieberman et al., 1972a; Crelin, Lieberman, and Klatt, forthcoming). I will, however, note some of the salient points in the discussion of particular fossils. I will first discuss the computer modeling technique that we employed to arrive at a functional assessment of these supralaryngeal vocal tracts. I think that it makes sense to approach the discussion of the reconstructions by first discussing the modeling technique. The modeling revealed that we really need to know only a few fairly gross aspects of the morphology of the supralaryngeal vocal tract to make meaningful statements about speech ability. The reason that this is so is itself one of the functional characteristics of human speech.

I'll begin the discussion of the modeling technique by returning to our studies of the speech capabilities of living nonhuman primates. This is a useful start since we can compare the results of our modeling with the actual phonomena. Figure 1 shows the left half of the head and neck of a young adult male chimpanzee sectioned in the midsagittal plane. Silicone rubber casts were made of the air passages, including the nasal cavity, by filling each side of the split air passages separately in the sectioned head and neck to insure perfect filling of the cavities. The casts from each side of a head and neck were then fused together to make a complete cast of the air passages. In Figure 2 the cast of the chimpanzee airways is shown together with casts made, following the same procedures, for newborn human and adult human. A cast of the reconstructed supralaryngeal airways of the La Chapelle-aux-Saints fossil also appears in this figure. In Figure 3 equal sized outlines of the air passages for these four vocal tracts are sketched.

Note the high position of the larynx in the newborn human and adult chimpanzee vocal tracts where the soft palate and epiglottis can be approximated. In the adult human vocal tract the soft palate and epiglottis are widely separated and cannot be approximated. The tongue is likewise at rest in newborn human and chimpanzee completely within the oral cavity, whereas in adult man the posterior third of the tongue is in a vertical position forming the anterior wall of the supralaryngeal pharyngeal cavity. Note, in particular, that there is practically no supralaryngeal portion of the pharynx present in the direct airway out from the larynx when the soft palate shuts off the nasal cavity in newborn human and in chimpanzee. In adult man half of the supralaryngeal vocal tract is formed by the pharyngeal cavity.

This difference between the chimpanzee and newborn supralaryngeal vocal tracts and that of adult Homo sapiens is a consequence of the opening of the larynx into the pharynx directly behind the oral cavity. In other words, the larynx opens almost directly into the oral cavity. This is the case for all living animals (Negus, 1949) with the exception of adult Homo sapiens. We really should use the term adultlike rather than adult since these differences appear to be fully developed by two years of age and are probably largely differentiated by six months of age (Lieberman, Harris, Wolf, and Russell, 1972b).

The functional distinctions that these anatomical differences confer on adult humans have been determined for respiration, swallowing, and the sense of smell. Kirchner (1970) notes that the respiratory efficiency of the adult human

Figure 1

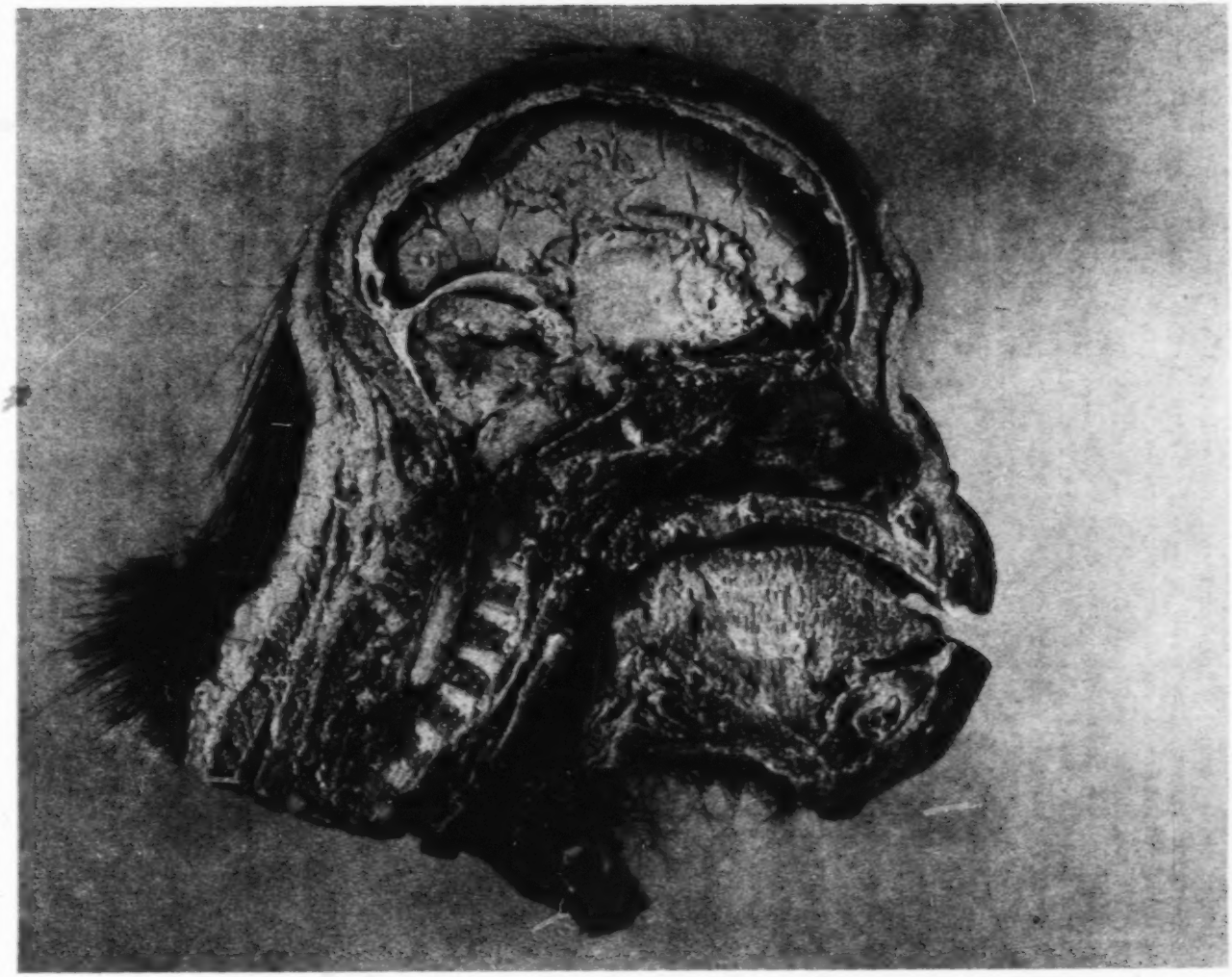


Figure 1: Left half of the head and neck of a young adult male chimpanzee sectioned in the midsagittal plane (after Lieberman et al., 1972).

Figure 2

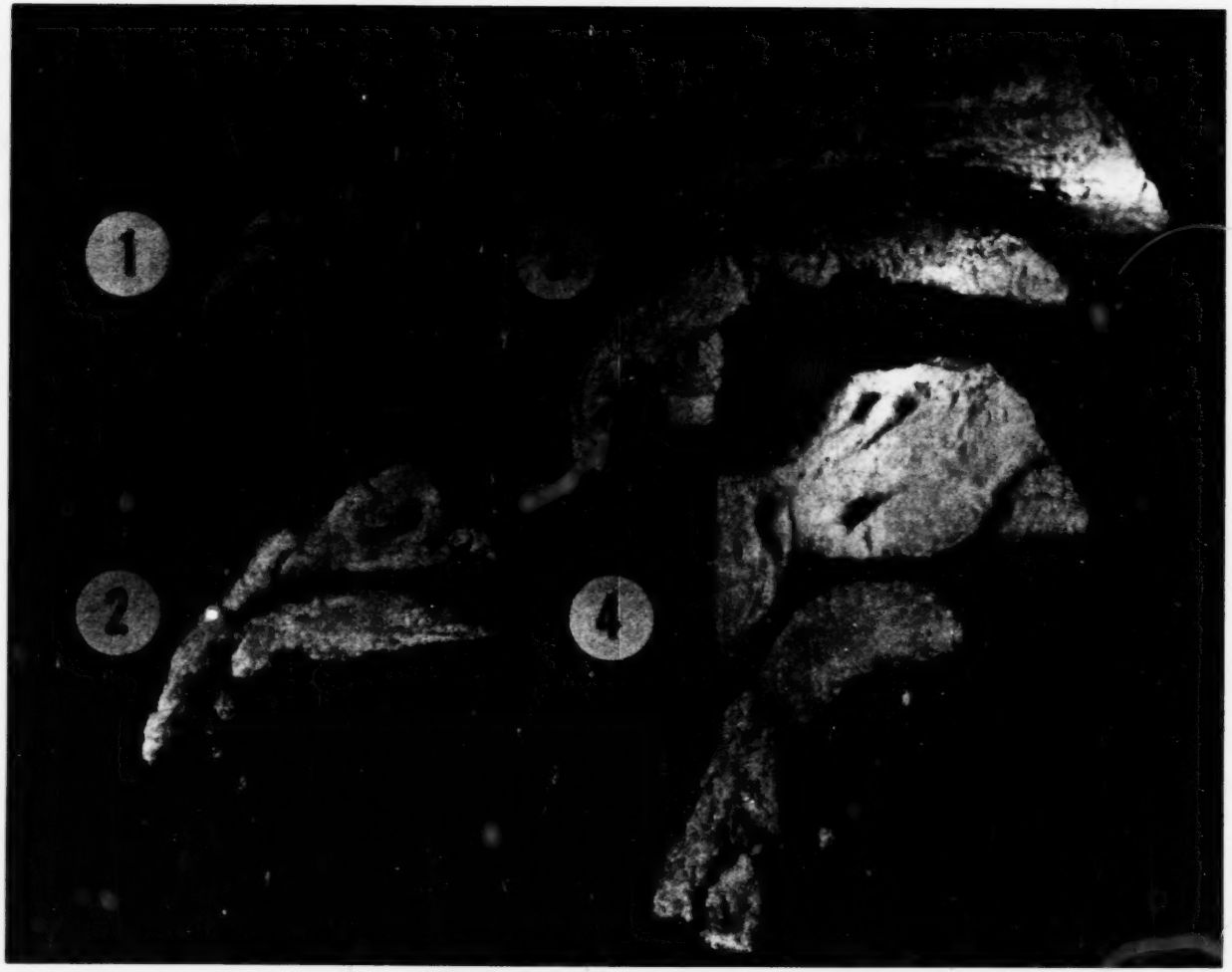


Figure 2: Casts of the nasal, oral, pharyngeal, and laryngeal cavities of (1) newborn Homo sapiens, (2) adult chimpanzee, (3) La Chapelle-aux-Saints reconstruction, and (4) adult Homo sapiens (after Lieberman et al., 1972).

Figure 3

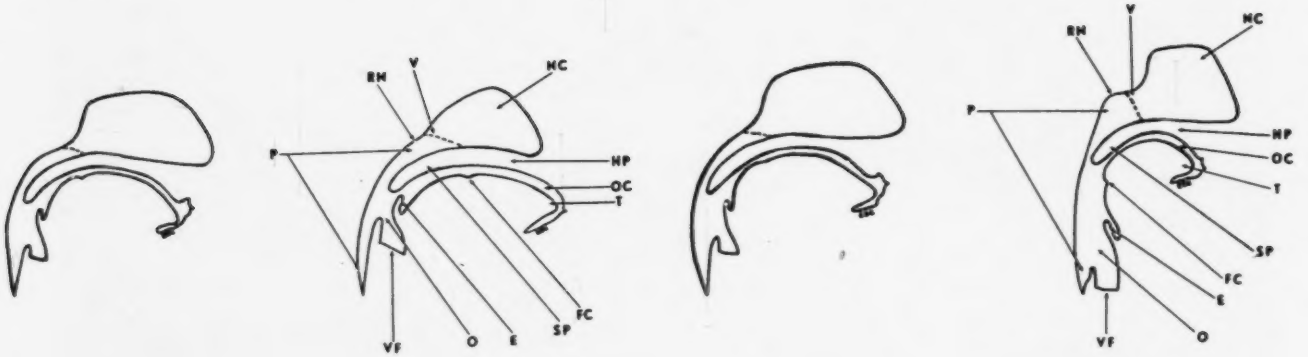


Figure 3: Diagrams of the air passages of (left to right) newborn human, adult chimpanzee, Neanderthal man, and adult human. The anatomical details that are keyed on the chimpanzee and adult man are as follows: P - Pharynx, RN - Roof of Nasopharynx, V - Vomer Bone, NC - Nasal Cavity, HP - Hard Palate, OC - Oral Cavity, T - Tongue, FC - Foramen Cecum, SP - Soft Palate, E - Epiglottis, O - Opening of Larynx into Pharynx, VF - Level of Vocal Folds (after Lieberman et al., 1972).

supralaryngeal airways is about half that of the newborn. The right-angle bend in the adult human supralaryngeal airway increases the flow resistance. The nonhuman supralaryngeal anatomy allows the oral cavity to be sealed from the rest of the airway during inspiration. This aids the sense of smell (Negus, 1949) and also allows an animal to breathe while its mouth contains a liquid (e.g., when a dog laps water). The adult human supralaryngeal airways also increase the possibility of asphyxiation. Food lodged in the pharynx can block the entrance to the larynx. This is not possible in nonhumans since the supralaryngeal pharynx serves as a pathway for both food and liquids and as an airway only in adult Homo sapiens.[1]

The functional distinctions that the differences in the anatomy of the supralaryngeal airways confer on speech production can be determined by modeling techniques. The source-filter theory of speech production, as I have noted before, states that speech is the result of the filtering action of the supralaryngeal vocal tract on the acoustic sources that excite it. Since the filtering properties of the filter are uniquely determined by the shape and length (the cross-sectional area function) of the supralaryngeal vocal tract it is possible to assess the properties of a particular vocal tract once we know the range of shapes that it can assume.

Note that this type of analysis will not tell us anything about the total range of phonetic variation. We would have to know the properties of the laryngeal source as well as the degree of motor control that a particular organism possessed. We can, however, assess the restraints that the supralaryngeal vocal tract itself imposes on the possible phonetic repertoire. The situation is similar to that which would occur if we found an ancient woodwind instrument made of brass. We would probably not be able to say very much about the reed, which would have decayed, but we would be able to determine some of the constraints that the instrument imposed on a performance. These constraints obviously would inherently structure the musical forms of the period. You can't write music that cannot be performed. We would not know all of the constraints, we could not say very much about the manual dexterity of the players or the general musical theory, but we would know more than would be the case if we had not found the ancient instrument.

We are in a somewhat better position when we study the reconstructed supralaryngeal vocal apparatus of an extinct hominid. We can tell something about the constraints on the phonetic repertoire. The interconnections that exist between the vocal apparatus and the perception of speech in Homo sapiens, however, allow us to make some more general inferences than would otherwise be the case.

---

[1] The human vocal tract is also inferior to the vocal tracts of hominids like La Chapelle-aux-Saints with respect to chewing. The reduction in the body of the mandible in modern Homo sapiens has reduced the tooth area. Dental studies have determined (Manly and Braley, 1950; Manly and Shiere, 1950; Manly and Vinton, 1951) that chewing efficiency in primates is solely a function of swept tooth area. Hominid forms that have smaller tooth areas have less efficient chewing. The reduction of the mandible in modern man therefore cannot be ascribed to enhancing chewing efficiency.

The technique that we have employed to assess the constraints imposed by the supralaryngeal vocal apparatus of an animal makes use of a computer model of the vocal tract. We really don't have to make use of this model. It would be possible, though somewhat tedious, to make actual models of possible supralaryngeal vocal tract configurations. If these models, made of plastic or metal, were excited by means of a rapid quasiperiodic series of puffs of air (i.e., an artificial larynx) we would be able to hear the actual vowel-like sounds that a particular vocal tract configuration produced. If we systematically made models that covered the range of possible vocal tract configurations we could determine the constraints that the supralaryngeal vocal tract morphology imposed, independent of the possible constraints determined by limitations on motor control, etc. We would be, of course, restricted to steady-state vowels since we could not rapidly change the shape of the vocal tract, but we could generalize our results to consonants since we could model the articulatory configurations that occur at the start and end of typical consonant-vowel sequences. Note that these modeling techniques allow us to assess the limits on the phonetic repertoire that follow from the anatomy of the supralaryngeal vocal tract, independent of muscular or neural control and independent of the dialect, habits, etc., of the animal whose vocal tract we would be modeling. The technology for making these mechanical models existed at the end of the eighteenth century. Von Kempelen's (1791) famous talking machine modeled the human vocal tract by mechanical means. The method that we have employed simply makes use of the technology of the third quarter of the twentieth century.

Chimpanzee, Newborn and Adult *Homo sapiens*

In Figure 4 three area functions are shown for the chimpanzee vocal tract, derived from the sectioned head and neck shown in Figure 1. The silicone rubber casting and schematic drawing of this vocal tract are shown in Figures 2 and 3 respectively. The area functions shown in Figure 4 represent the best approximations that we could get to the human vowels [a], [i], and [u]. We systematically drew area functions on an oscilloscope input to a PDP 9 computer with a light pen. The computer had been programmed to calculate the formant frequencies that corresponded to these area functions. The details of the computer program are discussed by Henke (1966). The computer allowed us conveniently and rapidly to make hundreds of possible supralaryngeal vocal tract models. We thus could explore the acoustic consequences of all possible chimpanzee supralaryngeal vocal tract configurations without waiting for a chimpanzee to actually produce these shapes. We used the same procedure to explore the possible range of supralaryngeal vocal tract shapes for the newborn human supralaryngeal vocal tract shown in Figures 2 and 3. We were guided in these simulations by the morphology of the head and neck, i.e., the relative thickness and position of the tongue, the lips, the velum, and the position of the pharynx relative to the larynx and oral cavity. We were also able to make use of cineradiographic pictures of newborn infants during cry and swallowing (Truby, Bosma, and Lind, 1965). The results of these simulations are shown in Figure 5. In Figure 5 the formant frequencies of the three area functions of Figure 4 are plotted, together with an additional data point (X) for human newborn. The loops labeled with phonetic symbols represent the data points for a sample of real utterances derived from 76 adult men, women, and adolescent children producing American-English vowels (Peterson and Barney, 1952). In Figure 6 we have reproduced the actual data points for this sample of real human vowels. Note that the chimpanzee and newborn human utterances cover only a small portion of the adult human "vowel space." In other

Figure 4

/i/			/ɑ/			/u/		
Formant	Freq.	Freq./1.7	Formant	Freq.	Freq./1.7	Formant	Freq.	Freq./1.7
1	610	360	1	1220	720	1	830	490
2	3400	2000	2	2550	1500	2	1800	1060
3	4420	2600	3	5070	2980	3	4080	2390

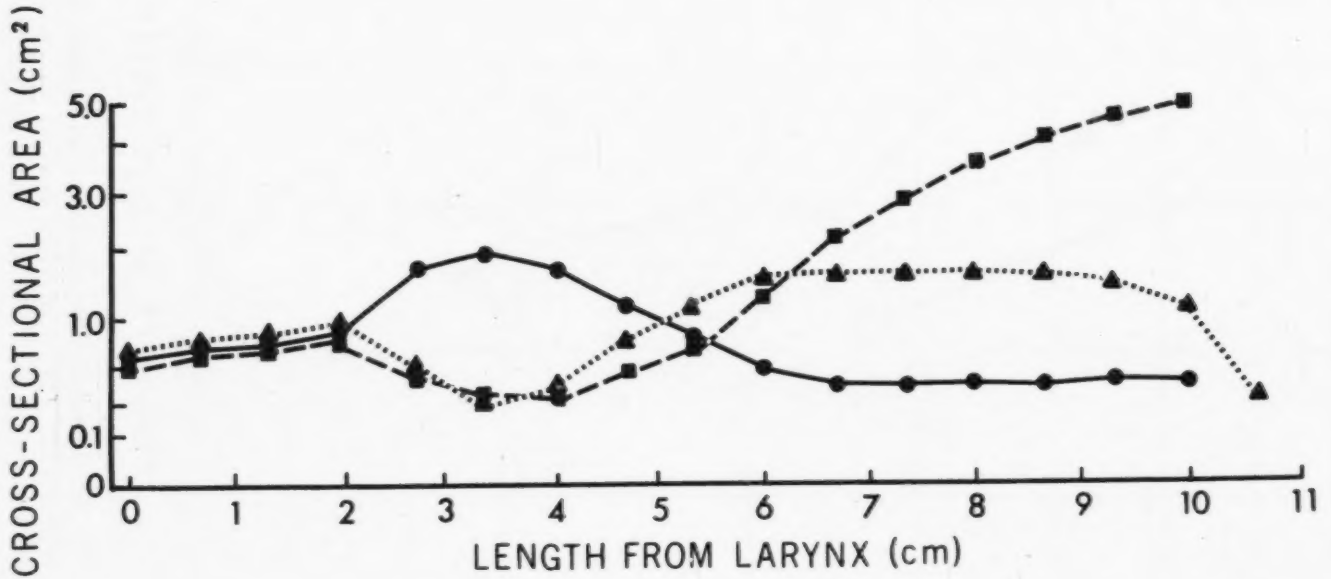


Figure 4: Chimpanzee supralaryngeal vocal tract area functions modeled on computer. These functions were the "best" approximations that could be produced, given the anatomic limitations of the chimpanzee, to the human vowels [i], [a], and [u]. The formant frequencies calculated by the computer program for each vowel are tabulated and scaled to the average dimensions of the adult human vocal tract (after Lieberman et al., 1972).

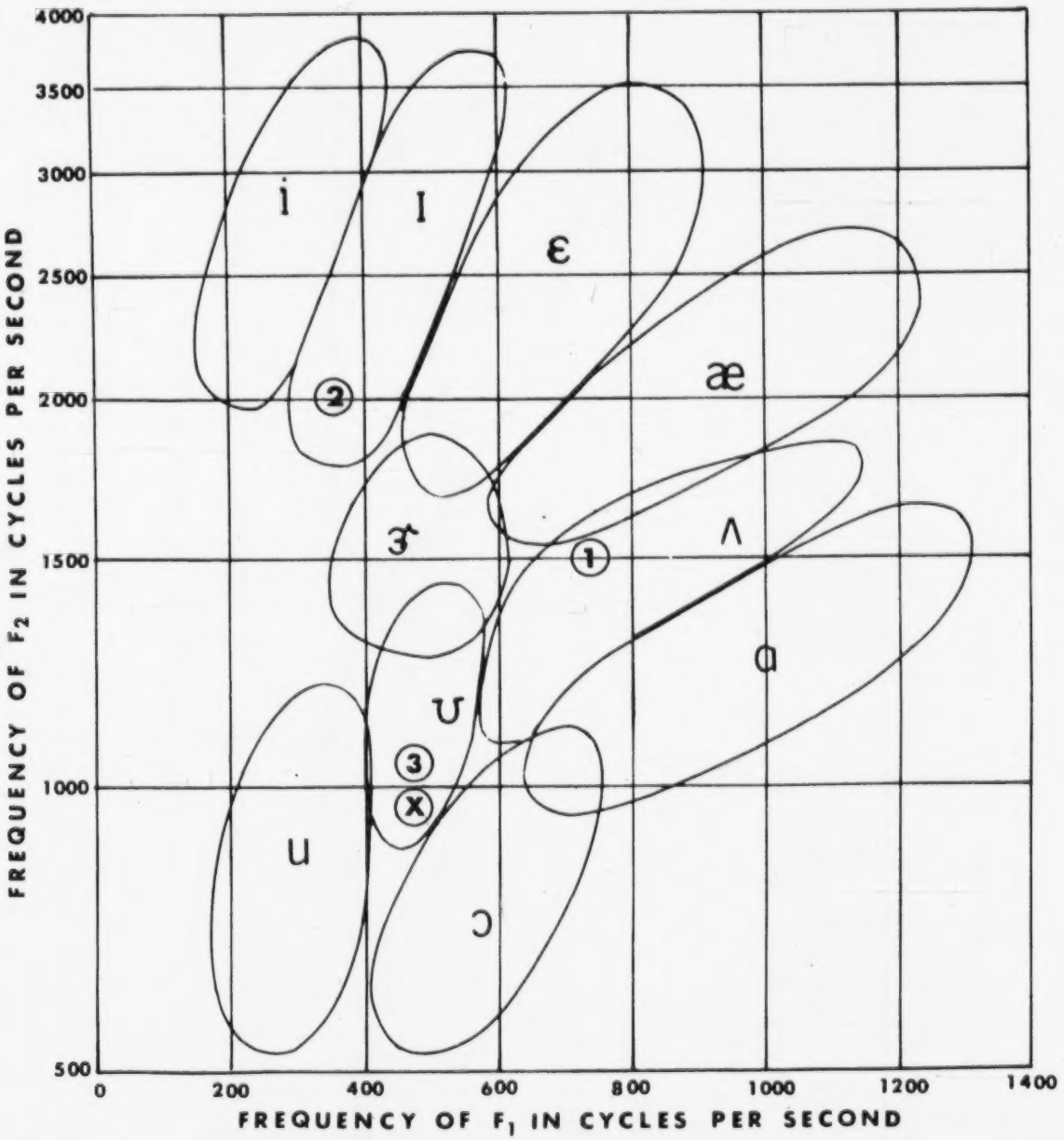


Figure 5: Plot of formant frequencies for chimpanzee vowels of Figure 4, data points (1), (2), and (3), scaled to correspond to the size of the adult human vocal tract. Data point (X) represents an additional point for human newborn. The closed loops enclose 90 percent of the data points derived from a sample of 76 adult men, women, and children producing American-English vowels (Peterson and Barney, 1952). Note that the chimpanzee and newborn vocal tracts cannot produce the vowels [i], [u], and [a] (after Lieberman et al., 1972).

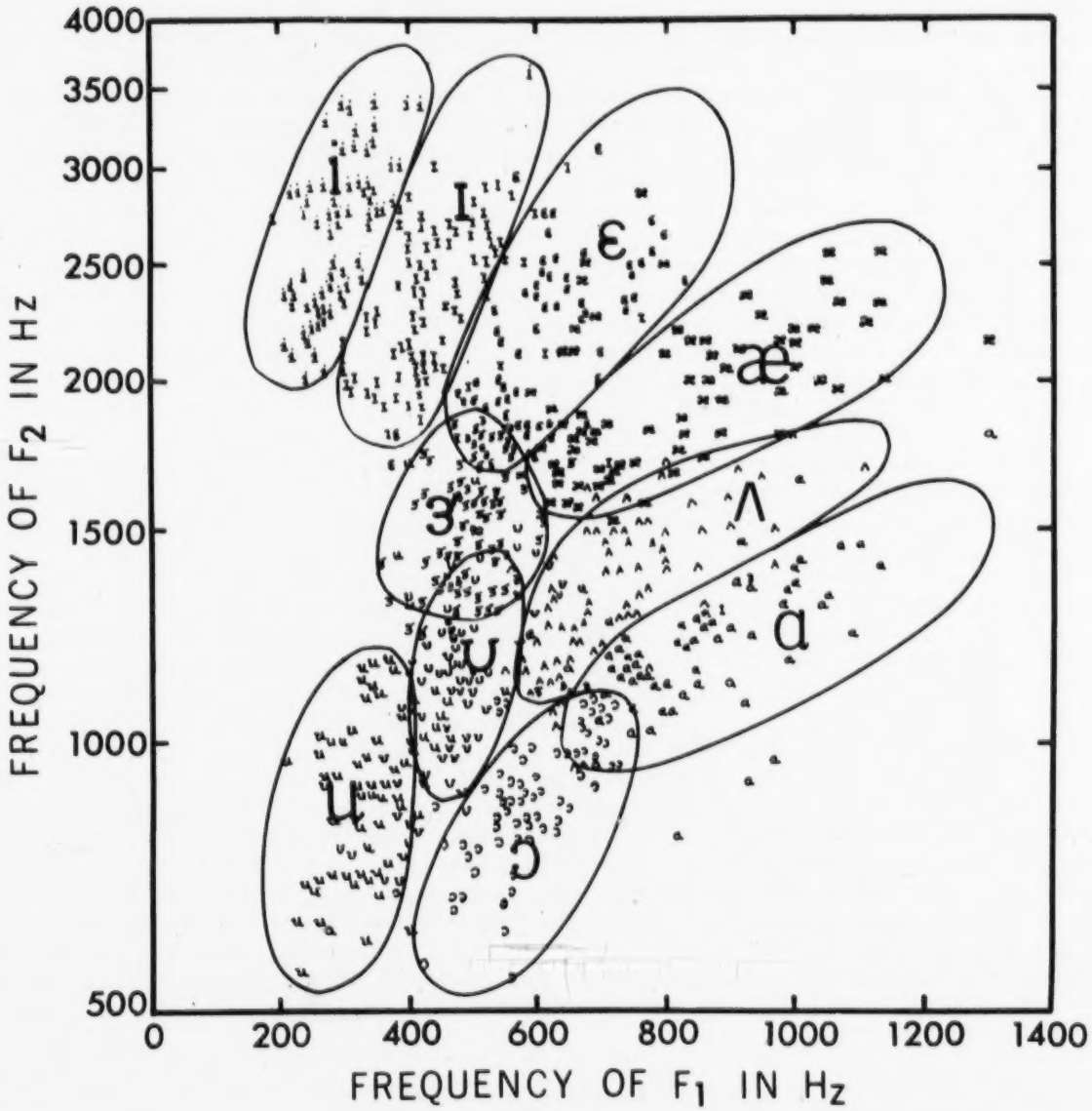


Figure 6: Formant frequencies of American-English vowels for a sample of 76 adult men, adult women, and children. The closed loops enclose 90 percent of the data points in each vowel category (after Peterson and Barney, 1952).

words, the chimpanzee and newborn vocal tracts according to this modeling technique inherently do not appear to be able to produce vowels like [a], [i], and [u].

All normal human speakers can inherently produce these vowels. Any human, if he is raised in an American-English environment will be able to produce these vowels. The modeling of the chimpanzee and newborn vocal tracts indicates that they could not, even if they had the requisite motor and neural abilities. The question that we are addressing is thus not whether chimpanzees and newborns can speak American English. It is whether they have the anatomical apparatus that would allow them to speak.

The results of the modeling technique can, of course, be checked against the actual utterances of chimpanzees and newborn Homo sapiens. When this is done it is evident that the actual vowels of newborn Homo sapiens agree with the computer simulation (Irwin, 1948; Lynip, 1951; Lieberman et al., 1972b). The chimpanzee simulation appears to encompass a greater range than has been observed so far in the acoustic analysis of chimpanzee vocalizations (Lieberman, 1968). This may merely indicate that the acoustic analyses so far derived from chimpanzees do not represent the total chimpanzee repertoire. It is, however, apparent that the computer simulation does not appear to be showing a smaller vowel space than is actually the case. The computer simulation for adult Homo sapiens corresponds with that observed (Chiba and Kajiyama, 1958; Fant, 1960; Peterson and Barney, 1952) and is not plotted here.

The vowel diagrams in Figures 5 and 6 are really an indirect way of showing that the chimpanzee and newborn cannot generate supralaryngeal vocal tract area functions like those shown in Figure 7. These three configurations are the limiting articulations of a vowel triangle that is language universal (Troubetzkoy, 1939). It is not a question of the chimpanzee and newborn not being able to produce American-English vowels. They could not produce the vowel range that is necessary for any other language of Homo sapiens. Particular modern languages may lack one of these articulations, but they always include at least one of these vowels and/or the glides [y] and [w] which are functionally equivalent to [i] and [u]. It is important to remember that we are discussing the phonetic level rather than the phonemic. Claims that a particular language, e.g., Kabardian (Kuipers, 1960), has only one centralized vowel generally concern the phonemic level, i.e., the claim is that a particular language does not differentiate words at the phonemic level through vowel contrasts. At the phonetic level these languages make use of vowels like [i], [u], and [a] though these vowels' occurrences are conditioned by other segments. It is also important to note that a vocal tract that cannot produce the area functions necessary for [i], [a], and [u] also cannot produce velar consonants like [g] and [k]. These consonants also involve discontinuities at the midpoint of the supralaryngeal vocal tract. Dental and bilabial consonants like [d], [t], [b], and [p] are, however, possible.

Figure 7 shows a midsagittal outline of the vocal tract for the vowels [i], [a], and [u], as well as the cross-sectional areas of the vocal tract (Fant, 1960) and the frequency domain transfer functions for these vowels (Gold and Rabiner, 1968). Ten to one discontinuities in the area function at the vocal tract's midpoint are necessary to produce these vowels. It is possible to generate these discontinuities with the "bent" adult human supralaryngeal vocal tract

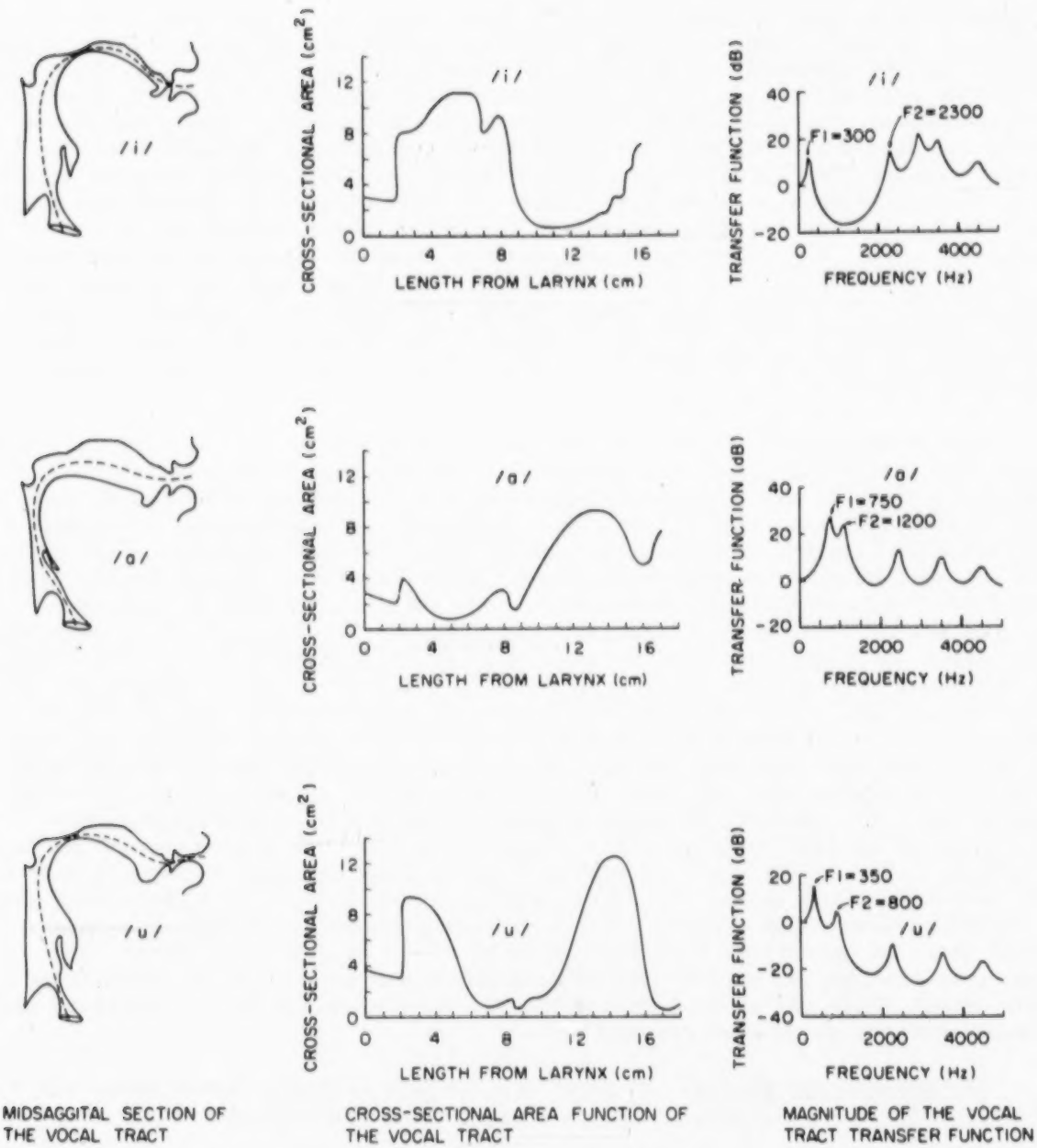


Figure 7: Illustrations of (left to right) approximate midsagittal sections, cross-sectional area functions, and acoustic transfer functions of the vocal tract for the vowels [i], [a], and [u] (after Lieberman et al., 1972).

since the cross-sectional areas of the oral and pharyngeal cavities can be independently manipulated in adult humans while a midpoint constriction is maintained. The supralaryngeal vocal tract in adult humans thus can, in effect, function as a "two tube" system. The lack of a supralaryngeal pharyngeal cavity in the direct airway from the larynx, at a right angle to the oral cavity, in chimpanzee and newborn humans restricts these forms to "single tube" resonant systems. In adult humans, muscles like the genioglossus can pull the pharyngeal portion of the tongue in an anterior direction, enlarging the pharyngeal cavity while the oral cavity is constricted, as in the production of [i]. In the production of [a], in adult humans, the pharyngeal constrictors reduce its cross-sectional area while the oral cavity is opened by lowering the mandible. It is impossible for the chimpanzee and newborn supralaryngeal vocal tracts to articulate these extreme discontinuities. They can only attempt to distort the tongue body in the oral cavity (see Figures 2, 3, and 4) to obtain changes in cross-sectional area. The intrinsic musculature and elastic properties of the tongue severely limit the range of deformations that the tongue body can be expected to employ. This is evident in cineradiographic observations of newborn cry and swallowing (Truby et al., 1965), of baboon cries (Zhinkin, 1963), and of the deformations of the oral and pharyngeal portions of the tongue in adult humans (Perkell, 1969).

Note that Figure 7 shows that the discontinuities in the [a], [i], and [u] area functions occur at or near the midpoint of the supralaryngeal vocal tract. Stevens (in press) has shown that the midpoint area discontinuity has an important functional value. It allows human speakers to produce signals that are acoustically distinct with relatively sloppy articulatory maneuvers. The first and second formant frequencies are maximally separated for [i], maximally centered for [a], and maximally lowered for [u]. When a human speaker wants to produce one of these vowels it is not necessary for him (or her) to be very precise about the position of the tongue. All that is necessary is an area function discontinuity within 1 cm or so from the midpoint. The formant frequencies will not vary perceptibly[2] (Flanagan, 1955) when the discontinuity shifts plus or minus 1 cm from the midpoint. This would not be the case for similar articulations if they were generated at any point other than the midpoint of the vocal tract. The vowels [a], [i], and [u] are thus optimal acoustic signals for communication. The speaker can produce maximally differentiated sounds without having to be terribly precise. All other vowels are both less distinct and less "stabile." The speaker must be more precise to produce acoustic signals that are not as distinct and separable. This factor is germane to one of the points that I raised earlier: how precise does the reconstruction of the supralaryngeal vocal tract of an extinct hominid have to be to yield meaningful data? The answer is that we can derive useful information without having to reconstruct fine detail since the crucial factor is essentially the ability to generate area discontinuities at or near the midpoint.

La Chapelle-aux-Saints. In Figures 2 and 3 a silicone rubber model and a sketch of the supralaryngeal vocal tract of the La Chapelle-aux-Saints Neanderthal

---

[2]Flanagan (1955) shows that human listeners are not able to discriminate stimuli that differ solely with respect to a single formant frequency unless the difference exceeds 60 Hz.

fossil are shown. It obviously was not possible to obtain this information directly from the soft tissue of this fossil hominid. The reconstruction of the supralaryngeal airways was effected by Edmund S. Crelin using the similarities that exist between this fossil and newborn human as a guide (Lieberman and Crelin, 1971; Lieberman et al., 1972a). The possible arthritic condition (Straus and Cave, 1957) of the La Chapelle-aux-Saints fossil has been raised in some criticisms of Crelin's reconstruction. Arthritic changes could no more have affected his supralaryngeal vocal tract than is the case in modern man.

Figure 8 shows a lateral view of the skull, vertebral column, and larynx of newborn and adult Homo sapiens and the reconstructed La Chapelle-aux-Saints fossil. Note that the geniohyoid muscle in adult Homo sapiens runs down and back from the symphysis of the mandible. This is necessarily the case because the hyoid bone is positioned below the mandible in adult Homo sapiens. The two anterior portions of the digastric muscles, which are not shown in Figure 8, also run down and back from the mandible for the same reason. When the facets into which these muscles are inserted at the symphysis of the mandible are examined, it is evident that the facets are likewise inclined to minimize the sheer forces for these muscles. The human chin appears to be a consequence of the inclination of these facets. The outwards inclination of the chin reflects the inclination of the inferior plane of the mandible at the symphysis. Muscles are essentially "glued" in place to their facets. Tubercles and fossae in this light may be simply regarded as adaptations that increase the strength of the muscle to bone bond by increasing the "glued" surface area. The inclination of the digastric and geniohyoid facets likewise serves to increase the functional strength of the muscle to bone bond by minimizing sheer forces. As Bernard Campbell (1966:2) succinctly notes, "Muscles leave marks where they are attached to bones, and from such marks we assess the form and size of the muscles." This is no less true for living than for extinct forms. When the corresponding features are examined in newborn Homo sapiens (Figure 8 and 9) it is evident that the nearly horizontal inclination of the facets of the geniohyoid and digastric muscles is a concomitant feature of the high position of the hyoid bone (Crelin, 1969:107-110). These muscles are nearly horizontal with respect to the symphysis of the mandible in newborn Homo sapiens. The facets therefore are nearly horizontal to minimize sheer forces. Newborn Homo sapiens thus lacks a chin.[3] When the mandible of the La Chapelle-aux-Saints fossil is examined, it is evident that the facets of these muscles resemble those of newborn Homo sapiens. The inclination of the styloid process away from the vertical plane is also similar in newborn Homo sapiens and the La Chapelle-aux-Saints fossil. When the base of the skull is examined (Figure 9) for newborn and adult Homo sapiens and the La Chapelle-aux-Saints fossil it is again apparent that the newborn Homo sapiens and fossil forms have many common features that differ from adult Homo sapiens. The sphenoid bone is, for example, exposed in newborn Homo sapiens and the La Chapelle-aux-Saints fossil between the vomer and the basilar part of the occipital. This is a skeletal feature that provides room for the larynx which is positioned high with respect to the mandible. There has to be room for the larynx behind the palate in newborn Homo sapiens and in the La-Chapelle-aux-Saints fossil. The qualitative

---

[3]The human chin is sometimes stated to be a reinforcement for the mandible. This is probably not the case. It more likely is a stress concentration point. It would be rather simple to resolve this point using the methods of stress analysis common in mechanical and civil engineering.

Figure 8

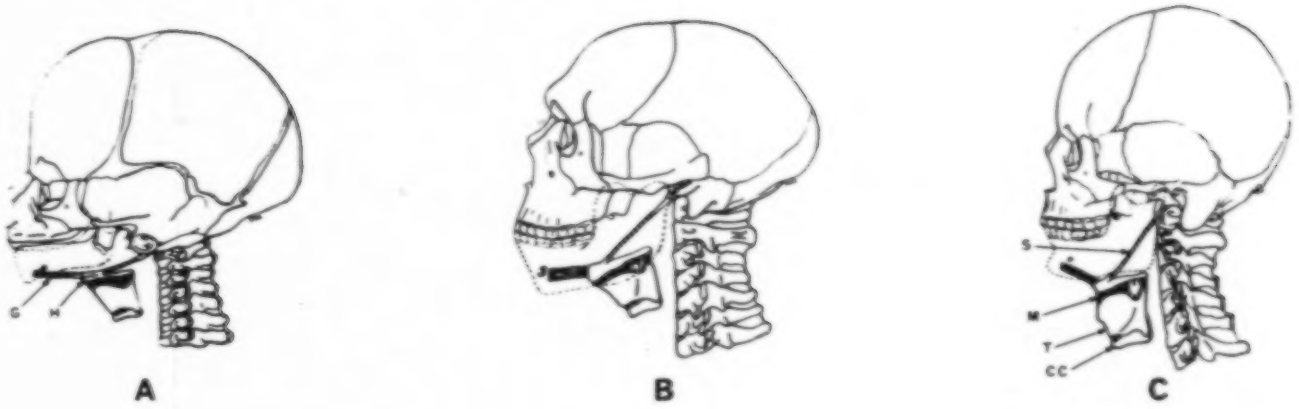


Figure 8: Skull, vertebral column, and larynx of Newborn (A), adult Man (C), and reconstruction of Neanderthal (B). G - Geniohyoid Muscle, H - Hyoid Bone, S - Stylohyoid Ligament, M - Thyrohyoid Membrane, T - Thyroid Cartilage, CC - Cricoid Cartilage. Note that the inclination of the styloid process away from the vertical plane in Newborn and Neanderthal results in a corresponding inclination in the stylohyoid ligament. The intersection of the stylohyoid ligament and geniohyoid muscle with the hyoid bone of the larynx occurs at a higher position in Newborn and Neanderthal. The high position of the larynx in the Neanderthal reconstruction follows, in part, from this intersection (after Lieberman and Crelin, 1971).

Figure 9

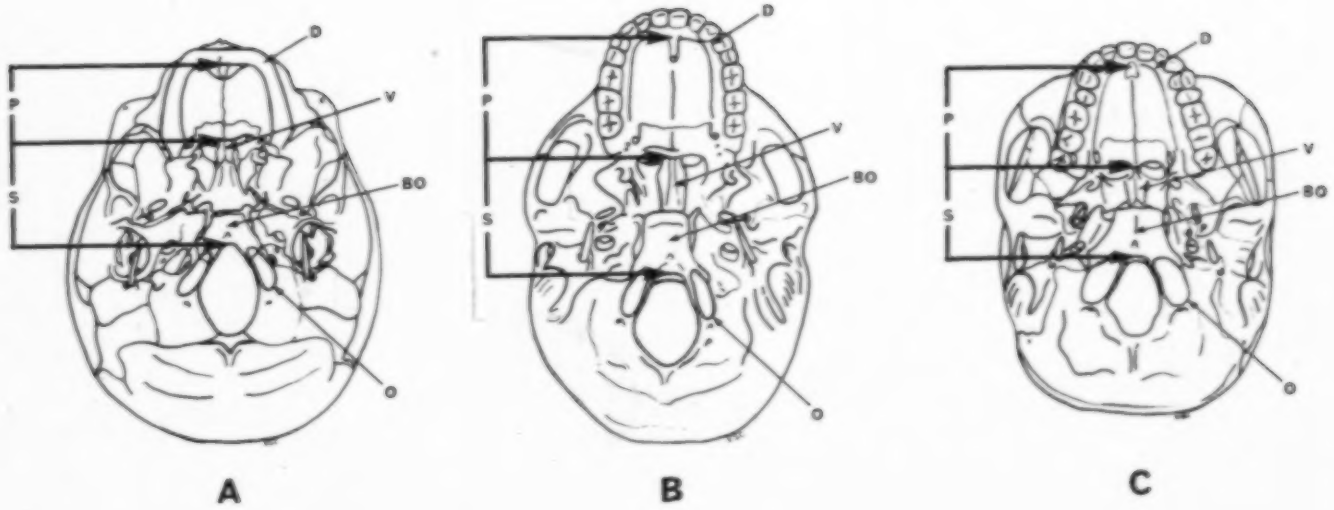


Figure 9: Inferior views of base of skull of Newborn (A), Neanderthal (B), and adult Man (C). D - Dental Arch, P - Palate, S - Distance Between Palate and Foramen Magnum, V - Vomer Bone, BO - Basilar Part of Occipital, O - Occipital Condyle (after Lieberman and Crelin, 1971).

difference in the morphology of the base of the skull, i.e., the exposure of the sphenoid, is a skeletal consequence of this anatomical necessity.

We do not claim that all the features of the La Chapelle-aux-Saints fossil are found in newborn Homo sapiens. This is definitely not the case. We are claiming that certain features, particularly those relating to the base of the skull and mandible, are similar. These similarities make possible a reasonably accurate reconstruction of the supralaryngeal vocal tract of the La Chapelle-aux-Saints fossil. Our observations are in accord with the results of Vlček's (1970) independent "Onto-phylogenetic" study of the development of a number of fossil skulls of Neanderthal infants. Vlček notes the presence of skeletal characteristics that are typical of both infant and adult Neanderthal fossils that are manifested during particular phases of the ontogenetic development of contemporary man. Other features that characterize adult Neanderthal man never appear in the ontogenetic development of contemporary man, while still other features that characterize contemporary man never are manifested in the fossil skulls. I will return to this data when I discuss the status of classic Neanderthal man. For the moment, it is relevant as an independent replication of the similarities between newborn Homo sapiens and the La Chapelle-aux-Saints fossil. Crelin's reconstruction of the supralaryngeal vocal tract of this fossil is also in accord with earlier attempts like that of Keith, which is discussed by Negus (1949), as well as the inferences of Coon (1966).

In Figure 10 the vowel space of the reconstructed La Chapelle-aux-Saints supralaryngeal vocal tract is presented. Each of the data points (N) represents attempts to produce vowels like [a], [i], or [u]. The labeled loops again refer to the Peterson-Barney (1952) data for actual human vowels. Note that the vowel space of the fossil is a subset of the human vowel space and that it is impossible to produce the "extreme" vowels [a], [i], and [u]. It is likewise impossible to produce the glides [y] or [w] or velar consonants like [g] and [k]. The Neanderthal supralaryngeal vocal tract also probably is not capable of making nasal versus nonnasal contrasts. Everything will tend to be nasalized. The modeled Neanderthal vowel space is probably too large since we allowed articulatory maneuvers that would have been rather acrobatic in modern man (Lieberman and Crelin, 1971). We tried to err on the side of making this fossil's phonetic ability more humanlike whenever we were in doubt.

Sterkfontein 5. In Figure 11 a silicone rubber model of the airways of the reconstructed supralaryngeal vocal tract of the Sterkfontein 5 cranium (Mrs. Ples) is shown together with the chimpanzee airways that appeared in Figure 2. Note the similarities. Crelin's reconstruction follows from the similarities that exist between this fossil and present-day orangutan, and to a lesser degree, chimpanzee. The reconstructed vocal tract has the same phonetic limitations as present-day ape's. The details of this reconstruction and the others that follow will be discussed in detail in a separate paper (Crelin et al., forthcoming).

Es-Skhūl V and Steinheim. In Figure 12 silicone rubber models of the reconstructed airways of the Es-Skhūl V and Steinheim fossils are shown together with the supralaryngeal airways of adult Homo sapiens. Note that the reconstructed supralaryngeal airways both have right-angle bends, that the pharyngeal cavity is part of the direct airway out of the larynx, that both resemble the supralaryngeal airways of adult modern man. The reconstructed Es-Skhūl V airway is completely modern. It would place no limits on its owner's phonetic repertoire

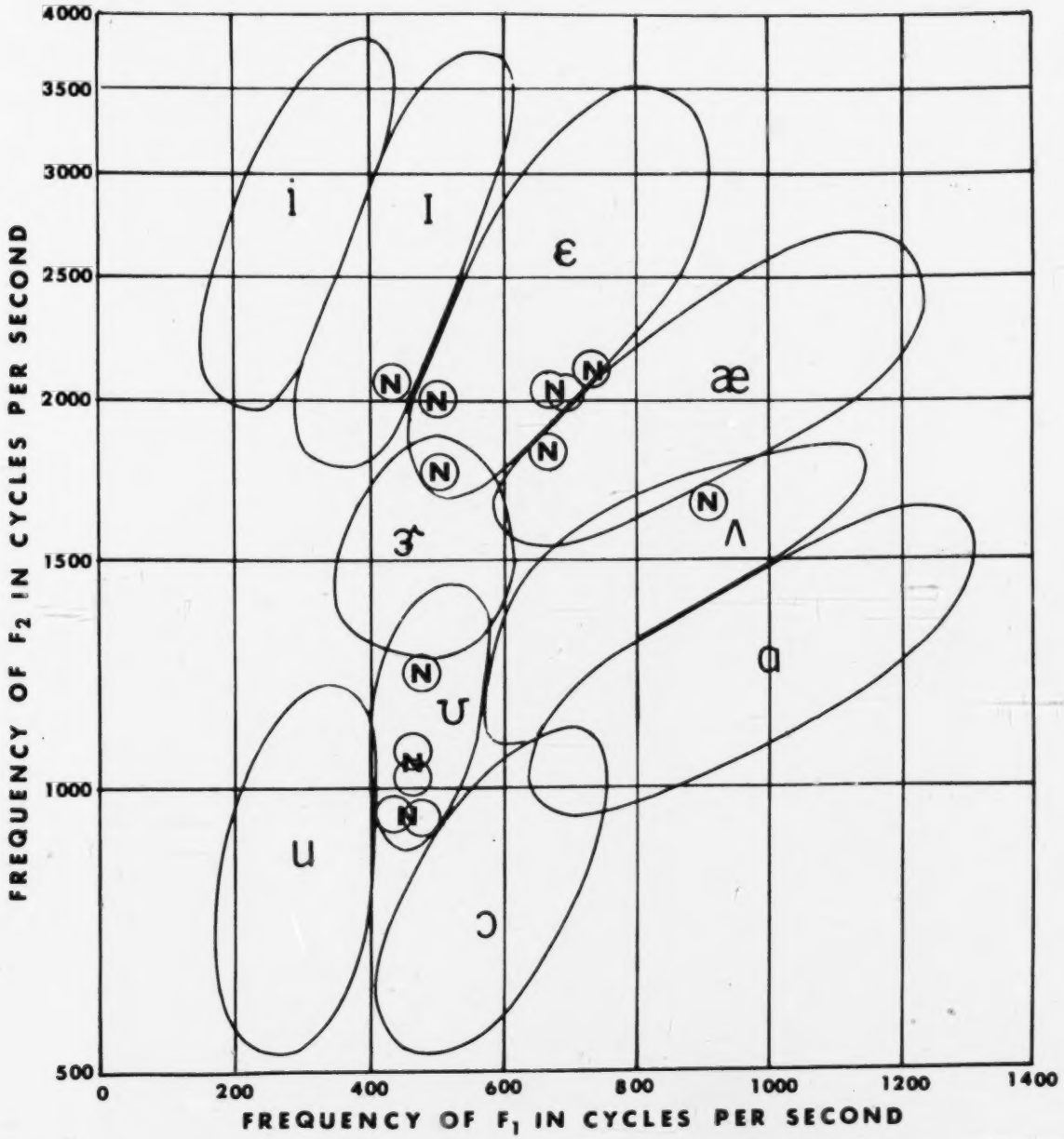


Figure 10: Plot of formant frequencies for reconstructed La Chapelle-aux-Saints supralaryngeal vocal tract in attempts to produce the vowels [i], [u], and [a]. Note that none of the data points (N) fall into the vowel loops that specify these vowels (after Lieberman and Crelin, 1971).

Figure 11

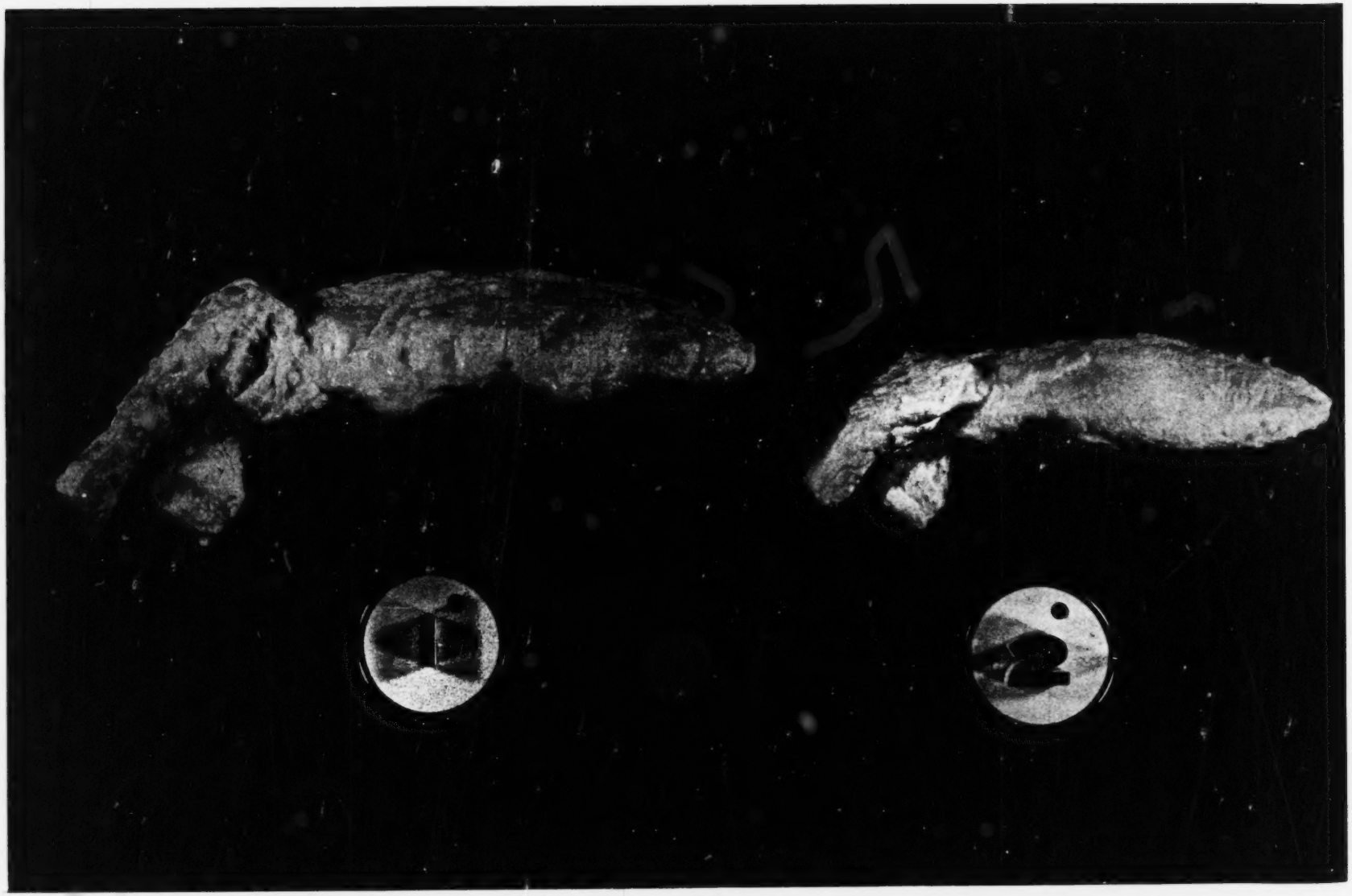


Figure 11: Casts of the oral, pharyngeal, and laryngeal cavities of (1) Sterkfontein 5 reconstruction and (2) chimpanzee. Note that the supralaryngeal airways of the Australopithecine fossil and chimpanzee are almost identical except for their size. The nasal cavities have been omitted to make the similarities in these "one tube" supralaryngeal vocal tracts more apparent.

Figure 12

Figure 12: Casts of the oral, pharyngeal, and laryngeal cavities of (1) Es-Skhūl V reconstruction, (2) Steinheim reconstruction, and (3) adult Homo sapiens (refer back to the key on Figure 3 for anatomical details). Note that the two fossil reconstructions' surralaryngeal airways both have a right angle bend and a pharyngeal cavity similar to that of modern Homo sapiens.

if he attempted to produce the full range of human speech. The Steinheim supralaryngeal airway, though it has some pongid features, is also functionally equivalent to a modern supralaryngeal vocal tract. It would have placed no restrictions on its owner's phonetic repertoire if he attempted to produce the full range of human speech.

Broken Hill (Rhodesian Man). In Figure 13 a silicone rubber model of the reconstructed supralaryngeal airways of Rhodesian man is shown together with a casting of the supralaryngeal airways of adult Homo sapiens. Note that despite the large oral cavity which follows from the large palate of this fossil, there is a right-angle bend in the supralaryngeal airway. This vocal tract appears to be an intermediate form. When it is modeled it can produce acoustic signals appropriate to the human vowels [a], [i], and [u] though supralaryngeal vocal tract configurations that are needed are not as stabile, i.e., resistant to articulatory sloppiness, as equivalent human vocal tract configurations. Note that the large palate in this fossil form occurs with a bent supralaryngeal vocal tract. The reduction of the palate in forms like Steinheim, Es-Skhūl V, and modern Homo sapiens therefore cannot be the factor that caused the larynx to descend.

SIGNIFICANCE OF RESULTS

Table 1 presents the results of the reconstructions and computer modeling so far discussed, together with the results that would be obtained for various fossils similar to the ones that we have examined. I have not attempted to list all the similar forms. Note that we have divided the table into two categories: fossil hominids with anatomical specializations necessary for human speech, and fossils lacking these specializations.

Neoteny

The first important point is that the anatomy necessary for producing the full range of sounds necessary for human speech represents a particular specialization that, at the present time, occurs only in normal adult Homo sapiens. It is clear that adult Homo sapiens does not particularly resemble newborn Homo sapiens.[4] In general, this is true of all primates (Schultz, 1968); the infantile forms of primates often do not resemble their adult forms. Schultz (1944, 1955), moreover, shows that the infantile forms of various nonhuman primates resemble newborn Homo sapiens, whereas the adult forms of these nonhuman primates diverge markedly from adult Homo sapiens. However, this does not mean that adult Homo sapiens has evolved by preserving neonatal features (Montagu, 1962), since it is apparent that modern man has his own unique specializations. The unique specializations of modern man include the anatomy necessary for the production of human speech. Table 1 shows that these specializations have evolved over at least the past 300,000 years and that until comparatively recent times, various types of hominids existed, including some who lacked the anatomical mechanisms necessary for articulate human speech.

---

[4] Benda (1969) shows that Down's Syndrome (mongolism) involves the retention of infantile morphology. Victims of this pathology, in some instances, retain the general proportions of the newborn skull. Their supralaryngeal vocal tracts retain the morphology of the newborn and they are unable to speak. They strikingly demonstrate that Homo sapiens has not evolved by retaining infantile characteristics.

Figure 13

Figure 13: Casts of the oral, pharyngeal, and laryngeal cavities of (1) Broken Hill reconstruction and (2) adult Homo sapiens (refer back to the key on Figure 3 for anatomical details). Note that the fossil reconstruction's supralaryngeal airway is an intermediate form. It has a right angle bend, but has a pharyngeal cavity smaller than that of adult, modern Homo sapiens.

TABLE 1

	- Human Supralaryngeal Vocal Tract	+ Human Supralaryngeal Vocal Tract
Australopithecines:	africanus robustus bosei	
Neanderthal	Saccopastore I	Steinheim
	Monte Circeo	
	Teschik-Tasch (infant)	
	La Ferrassie I	
	La Chapelle-aux-Saints	
	La Quina (infant)	Broken Hill
	Pech-de-l'Azé	
		Es-Skhūl V
		Djebel Kafzeh
	Solo 11	
	Shanidar I	Cro-Magnon
		modern Homo sapiens

The "Neanderthal Problem"

Note that Table 1 places a number of fossil forms that lacked speech into a category labeled "classic Neanderthal." A view that has enjoyed some popularity in recent years is that Neanderthal fossils do not substantially differ from modern Homo sapiens, that they simply form a subset of hominids with characteristics that grade imperceptibly from those typical of the modern population of Homo sapiens. An extreme formulation of this view is, for example, that "...no single measurement or even set of measurements can set Neanderthals apart from modern man" (Nett, in press); in other words, that Neanderthal man cannot be regarded as a separate species or even a separate variety distinct from Homo sapiens. This claim can be substantiated only if one includes fossils like Steinheim and Es-Skhūl V in the same class as forms like La Chapelle-aux-Saints. Quantitative multivariate analysis like that of Howells (1968) demonstrates that fossils like La Chapelle and La Ferrasie form a class that is quite distinct from modern man. The measurements contained in Patte's (1955) comprehensive work, as well as the observations of Vlček (1970) on the ontogenetic development of Neanderthal infants, indicate that this class of fossils, classic Neanderthal man, represents a specialization that diverged from the line (or lines) of more direct ancestors of Homo sapiens. Fossils like Steinheim and Es-Skhūl V, which are sometimes categorized as "generalized" Neanderthal, are functionally distinct from classic Neanderthal. These fossils exhibit the anatomical specializations necessary for human speech.

A general overlap between modern man and Neanderthal man is possible only if forms like Steinheim and Es-Skhūl V are put into the same class as La Chapelle, La Ferrasie, Monte Circeo, etc. Hominids who could have produced human speech would have to be classified with hominids who could not have produced human speech. This would be equivalent to putting forms that had the anatomical prerequisites for bipedal posture into the same class as forms that lacked this ability.

The question immediately arises: is this category, i.e., set of fossils labeled "classic Neanderthal," a separate species? It is useful to remember Darwin's definition of the term species. Darwin (1859:52) viewed the term species "...as one arbitrarily given for the sake of convenience to a set of individuals closely resembling each other, and that it does not essentially differ from the term variety, which is given to less distinct and more fluctuating forms." Darwin later notes (1859:485), "...the only distinction between species and well-marked varieties is that the latter are known, or believed, to be connected at the present day by intermediate gradations, whereas species were formerly thus connected." It is evident that intermediate fossil forms like Broken Hill man bridge the gap between classic Neanderthal man and modern Homo sapiens. We do not know, and we probably never will be able to know, all the traits that may have differentiated various hominid populations that are now extinct. We do not, for example, know whether viable progeny would have resulted from the mating of forms like Cro Magnon and La Quina. Even if we did know that viable progeny would result from the mating of classic Neanderthal and early Homo sapiens populations, we would not necessarily conclude that these forms were members of the same species. The term species as Darwin noted is simply a labeling device. Canus lupus and Canus familiarus are considered to be separate species even though they may freely mate and have viable progeny. The behavioral attributes of wolves and dogs make it important for people, for example, shepherds,

to place these animals into different species, even though some dogs, e.g., chihuahuas and St. Bernards, are more distinct morphologically and behaviorly and can't mate. The question of separate species labels for classic Neanderthal and other fossil hominid populations is thus probably an overworked question. We simply can note that different types of hominids apparently coexisted until comparatively recent times and that some of these hominids do not appear to have contributed to the present human gene pool.

Table 1 does have some bearing on the apparent absence of the specialization typical of classic Neanderthal man (e.g., La Chapelle, La Ferrasie, etc.) in modern man. Animal studies (Capranica, 1965) have established the role of vocalizations in courtship and mating. The presence or absence of humanlike speech probably would have served as a powerful factor in assortative mating. In the present population of modern man it is evident that linguistic differences and affinities play a powerful role in mate selection. We would expect this phenomenon to be accentuated when different hominid populations inherently were unable to produce the sounds of other groups. Sexual selection determined by speech patterns may thus have played a significant role in the divergence of groups like classic Neanderthal in Western Europe and the ancestral forms of modern Homo sapiens.

The Evolutionary Sequence

There is, unfortunately, a large gap in Table 1 since we have not yet been able to examine specimens of Homo erectus that have intact skull bases. It is, however, likely that the situation that typifies later hominid forms will also characterize Homo erectus--that Homo erectus most probably will not have the anatomy necessary for the production of the full range of human speech. Some forms, however, will undoubtedly be found that either had the necessary anatomy or that were intermediate forms. Evolution goes in small steps, and forms intermediate between Steinheim and the Australopithecines must have existed. We still are, like Darwin, at the mercy of the "Imperfection of the Geological Record."

We can, despite this gap, draw several inferences from Table 1. I would like to propose the following evolutionary sequence. The first phase of the evolution of human language must have relied on a system of gestures, facial expression, and vocal signals like those of present-day apes to communicate the semantic, i.e., cognitive, aspects of language. The Australopithecines must have had cognitive abilities surpassing those of present-day apes. Even early Australopithecines had cognitive abilities that surpassed the levels of present-day apes; late forms would have developed superior abilities as evolution continued step by step and mutations favoring larger relative brain sizes were retained. The retention of mutations leading to larger relative brain sizes is itself a sign that cognitive ability had a selective advantage. We can reasonably infer that activities like tool making and collective social enterprises like hunting were important attributes of Australopithecine culture.

Although the vocal apparatus of forms like Australopithecus africanus does not appear to differ significantly from those of present-day apes, vocal communications undoubtedly played a part in their linguistic system. Our reconstructions can tell us nothing about the larynx; however, it is almost certain that the laryngeal mechanisms of these forms was at least as developed as those of present-day apes. As Negus (1949) observed, there is a continual elaboration of

the larynx as we ascend the phylogenetic scale in terrestrial animals. The larynges of animals like wolves are capable of producing a number of distinct calls that serve as vehicles of vocal communication. The same is true for the larynges of chimpanzees and gorillas. Studies like that of Kelemen (1948), which have attempted to show that chimpanzees cannot talk because of laryngeal deficiencies, are not correct. Kelemen shows that the chimpanzee's larynx is different from the larynx of a normal adult human male. The chimpanzee's larynx will not produce the range of fundamental frequencies typical of adult human males; however, it can produce a variety of sound contrasts. Many of these sound contrasts indeed occur in human languages. A present-day chimpanzee, if it made maximum use of its larynx and supralaryngeal vocal tract, could, for example, produce the following sound contrasts:

a. Voiced versus Unvoiced, i.e., excitation of the vocal tract by the quasi-periodic output of the larynx versus turbulent noise excitation generated by opening the larynx slightly and expelling air at a high flow rate.

b. High Fundamental versus Normal Fundamental, i.e., adjusting the larynx so phonation occurs in the falsetto register rather than the modal chest register (van den Berg, 1960). The larynx has several modes of phonation which result in acoustic signals that are quite distinct. In falsetto the fundamental frequency is high and the glottal source's energy spectrum has comparatively little energy at its higher harmonics.

c. Low Fundamental versus Normal Fundamental, i.e., adjusting the larynx to a lower register. This lower register termed "fry" produces very low fundamental frequencies (Hollien, Moore, Wendahl, and Michel, 1966) that are irregular (Lieberman, 1963).

d. Dynamic Fundamental Frequency Variations, e.g., low to high, high to low. Variations like these occur in many human tone languages.

e. Strident High Energy Laryngeal Excitation, i.e., the high fundamental frequency, breathy output that can be observed in some chimpanzee vocalizations (Lieberman, 1968) as well as in the cries of human newborn (Lieberman et al, 1972b).

f. Continuent versus Interrupted, i.e., the temporal pattern of laryngeal excitation can be varied. This can be observed in the calls of present-day monkeys and apes (Lieberman, 1968).

g. Oral versus Nonoral, i.e., the animal can produce a call with his oral cavity sealed or with his oral cavity open. This can be observed in present-day gorilla where the low energy, low fundamental frequency sounds that sometimes accompany feeding appear to be produced with the oral cavity sealed by the epiglottis (Lieberman, 1968).

h. Lip Rounding and Laryngeal Lowering. Chimpanzees have the anatomic ability of rounding their lips and/or lowering their larynges while they produce a call. Both of these articulatory gestures could produce a formant frequency pattern that had falling transitions.

i. Flared Lips and Laryngeal Raising. Chimpanzees could either flare their lips and/or raise their larynges while they produced a call. This would generate a rising formant frequency pattern.

j. Bilabial Closures and Releases. Sounds like [b] and [p], as well as prevoiced [b] [like that occurring in Spanish, for example (Lisker and Abramson, 1964)], could be produced by controlling the timing between the opening and closing of the larynx and the lips.

k. Dental Closures and Releases. Sounds like [d] and [t] (Lisker and Abramson, 1964) could be produced by varying the timing between a closure effected by the tongue blade against the alveolar ridge and the opening and closing of the larynx.

Australopithecines could have generated all of the above sound contrasts if they had the requisite motor control and the neural ability to perceive the differences in sound quality that are the consequences of these articulatory maneuvers. Most of these phonetic contrasts, i.e., "features" (Jacobson, Fant, and Halle, 1952) have been observed in the vocal communications of present-day nonhuman primates. Present-day human languages make use of all of these sound contrasts. The combination of articulatory features like (j) and (k) and timing features like (f) could also generate sounds like [f], [v], [s], etc. It is quite probable that late Australopithecines and various forms of Homo erectus made use of these sound contrasts to communicate. The transference of patterns of automatized behavior (discussed early in this paper) from activities like tool making and hunting would have facilitated the acquisition of the motor skills necessary to produce these sounds. The role of hunting would have placed a premimum on communication out of the line of sight, communication that, furthermore, left the hunter's hand free.

The neural mechanisms necessary for the differentiation of these sounds appear to exist in present-day primates. Wollberg and Newman (1972), for example, have shown that squirrel monkeys (Saimiri sciureus) possess auditory receptors "tuned" to one of the vocal calls that these monkeys make use of in their communications. Similar results have been demonstrated for frogs (Capranica, 1965). Although gestural communication (Hewes, 1971) undoubtedly played a more important role in the communications of these early hominids than for modern man, I think that it is most unlikely that vocal communications also did not play an important role.

The crucial stage in the evolution of human language would appear to be the development of the bent supralaryngeal vocal tract of modern man. Table 1 shows a divergence in the paths of evolution. Some hominids appear to have retained the communications system typical of the Australopithecines, a mixed system that relied on both gestural and vocal components. Other hominids appear to have followed an evolutionary path resulting in almost total dependence on the vocal component for language and relegating the gestural component to a secondary, paralinguistic function. The process would have been gradual, following from the prior existence of vocal signals in linguistic communication.

As I have noted before, the bent supralaryngeal vocal tract that appears in forms like present-day Homo sapiens, and the Steinheim, Es-Skhūl V, and Broken Hill fossils, allows its possessors to generate acoustic signals that have very

distinct acoustic properties and that are very easy to produce. These signals are, in a sense, optimal acoustic signals (Lieberman, 1970). If vocal communications were already part of the linguistic system of early hominids, then mutations that extended the range and the efficiency of the signaling process would have been retained in forms like Steinheim. The bent supralaryngeal vocal tract is otherwise a burden for basic vegetative functions. It would not have been retained unless it had conferred an adaptive advantage. The initial adaptive value of the bent supralaryngeal vocal tract would have been its value in increasing the inventory of vocal signals and, moreover, in providing more efficient vocal signals.

The neural mechanisms necessary to perceive these new signals would, in all likelihood, have been available to hominids like Steinheim. Recent electrophysiological data (Miller, Sutton, Pfingst, Ryan, and Beaton, 1972) shows that animals like rhesus monkey, Maccaa mullata, will develop neural detectors which identify signals important to the animal. Receptors in the auditory cortex responsive to a 200 Hz sinusoid were discovered after the animals were trained by the classic methods of conditioning to respond behaviorly to this acoustic signal. These neural detectors could not be found in the auditory cortex of untrained animals. The auditory system of these primates thus appears to be "plastic." Receptive neural devices can be formed to respond to acoustic signals that the animal finds useful. These results are in accord with behavioral experiments involving human subjects where "categorical" responses to arbitrary auditory signals can be produced by means of operant conditioning techniques (Lane, 1965). They are also in accord with the results of classic conditioning experiments like those reported by Pavlov. The dogs learned to identify and to respond decisively to the sound of a bell, an "unnatural" sound for a dog. The dog obviously had to learn to identify the bell. Hominids like Steinheim who had the potential to make new acoustic signals would also have had the ability to learn to respond to these sounds in an automatized way. The plasticity of the primate auditory system would have provided the initial mechanism for learning these new sounds.

Later stages in the evolution of human language probably involved the retention of mutations that had innately determined neural mechanisms that were tuned to these new sounds. By innately determined, I do not mean that the organism needs no interaction with the environment to learn to perceive these sounds. The evidence instead suggests that humans are innately predisposed to learn to respond to the sounds of speech. Experiments with 6-week old infants (Eimas, Siqueland, Jusczyk, and Vigorito, 1971; Morse, 1971) show that they respond to the acoustic cues that differentiate sounds like [b] and [p] in the same manner as adults. These acoustic distinctions involve 10 msec differences in the timing of the delay between the start of the acoustic signal that occurs when a human speaker opens his lips and the start of phonation. It is most improbable that 6-week old infants could learn to respond to these signals unless there was some innate predisposition for this sound contrast to be perceived. This surely is not surprising. Human infants really do not learn the complex physiologic maneuvers associated with normal respiration. They have built-in knowledge. The case for the neural mechanisms involved in the perception of human speech is not as simple as that for respiration. Some contact with a speech environment is necessary. Deaf children, for example, though they at first produce the vocalizations of normal children, become quiet after six months of age (Lenneburg, 1967). Nottebohm (1970) shows similar effects in birds. Some aspects of the

bird's vocal behavior are manifested even when the bird is raised in isolation. Other important aspects of the bird's vocal behavior develop only when the bird is exposed to a normal communicative environment.

At some late stage, that is, late with respect to the initial evolution of the bent supralaryngeal vocal tract, the neural mechanisms that are necessary for the process of speech encoding would have evolved. The humanlike supralaryngeal vocal tract would have been retained initially for the acoustically distinct and articulatory facile signals that it could generate. The acoustic properties of sounds like the vowels [a], [i], and [u] and the glides [y] and [w], which allow a listener to determine the size of the speaker's supralaryngeal vocal tract, would have preadapted the communications system for speech encoding.

When a human listener hears a sound like the word bat, as it is produced by an intermediate sized supralaryngeal vocal tract, it is indeterminate. Ladefoged and Broadbent (1957), for example, show that a listener will perceive this sound as the word bit if he is led to believe that it was produced by a long vocal tract. The same listener will perceive the same sound as but if he is led to believe that it was produced by a small, i.e., short vocal tract. The listener, in effect, normalizes the signal to take account of the acoustic properties of different sized vocal tracts. The listener responds as though he is interpreting the acoustic signal in terms of the articulatory gestures that a speaker would employ to generate the word. The perception of human speech is generally structured in terms of the articulatory gestures that underlie the acoustic signal (Liberman et al., 1967). This process, as I noted earlier, is the basis of the encoding which allows human speech to transmit information at the rate of 20 to 30 segments per second. Signals like the vowels [a], [u], and [i] and the glides [y] and [w] are determinate in the sense that a particular formant pattern could have been generated by means of only one vocal tract using a particular articulatory maneuver (Stevens and House, 1955; Lindblom and Sundberg, 1969). A listener can use these vowels to identify instantly the size of the supralaryngeal vocal tract that he (or she) is listening to (Darwin, 1971; Rand, 1971). These vowels can indeed serve the same function in the recognition of human speech by computer. Gerstman (1967), for example, derives the size of a particular speaker's vocal tract from these vowels to identify the speaker's other vowels. Without this information it is impossible to assign a particular acoustic signal into the correct vowel class. The computer, like a human, has to know the size of the speaker's supralaryngeal vocal tract. The process of speech encoding need not have followed the exact path that I have proposed. Other sounds, like [s] can provide a listener (or a computer) with information about the size of the speaker's vocal tract. As I noted before, the Australopithecines had the anatomical prerequisites for producing sounds like [s], so the process of speech encoding and the evolution of the human supralaryngeal vocal tract may have been coeval from the start. It is clear, however, that evolution goes by small steps and what we have in present-day man is a fully encoded speech system with a speech producing anatomy that is highly adapted to this function. Other, now extinct, hominids like classic Neanderthal man had speech producing anatomy that clearly was not as well adapted for speech encoding. It is, therefore, reasonable to conclude that speech encoding either was more rudimentary or not present.

However, it is important to conclude with the point that language does not necessarily have to involve the process of speech encoding and rapid information transfer. The remains of Neanderthal culture all point to the presence of

linguistic ability. Conversely, birds may have the potential for rapid information transfer (Greenewalt, 1967); however, birds lack the cognitive ability that is also a necessary factor in language. It is most unlikely that birds could develop a complex language unless they also had larger brains.

Human language is the result of the convergence of many factors: automatization, cognitive ability, and speech encoding. The particular form that human language has taken, however, appears to be the result of the evolution of the human supralaryngeal vocal apparatus. The supralaryngeal vocal apparatus that differentiates present-day Homo sapiens from all living animals thus is as important a factor in the late stage of hominid evolution as dentition and bipedal posture were in earlier stages.

REFERENCES

Benda, C. E. (1969) Down's Syndrome, Mongolism and Its Management. (New York: Grune and Stratton).

Campbell, B. (1966) Human Evolution, an Introduction to Man's Adaptations. (Chicago: Aldine).

Capranica, R. R. (1965) The Evoked Vocal Response of the Bullfrog. (Cambridge, Mass.: MIT Press).

Chiba, T. and M. Kajiyama. (1958) The Vowel, Its Nature and Structure. (Tokyo: Phonetic Society of Japan).

Coon, C. S. (1966) The Origin of Races. (New York: Knopf).

Crelin, E. W. (1969) Anatomy of the Newborn: An Atlas. (Phildelphia: Lea and Febiger).

Crelin, E. W., P. Lieberman, and D. H. Klatt. (forthcoming) Anatomy and related phonetic ability of the Skhūl V, Steinheim, and Rhodesian Fossils and the Pleisianthropus reconstruction.

Darwin, C. (1859) On the Origin of Species, Facsimile edition. (New York: Atheneum).

Darwin, C. J. (1971) Ear differences in the recall of fricatives and vowels. Quart. J. Exp. Psychol. 23, 386-392.

Eimas, P. D., E. R. Siqueland, P. Jusczyk, and J. Vigorito. (1971) Speech Perception in infants. Science 171, 303-306.

Fant, G. (1960) Acoustic Theory of Speech Production. (The Hague: Mouton).

Flanagan, J. L. (1955) A difference limen for vowel formant frequency. J. Acoust. Soc. Amer. 27, 613-617.

Gardner, R. A. and B. T. Gardner. (1969) Teaching sign language to a chimpanzee. Science 165, 664-672.

Gerstman, L. (1967) Classification of self-normalized vowels. Proceedings of IEEE Conference on Speech Communication and Processing 97-100. (Bedford, Mass.: Air Force Cambridge Research Labs).

Gold, B. and L. R. Rabiner. (1968) Analysis of digital and analog formant synthesizers. IEEE Trans. Audio Electroacoustics AU-16, 81-94.

Goodall, J. v. L. (1971) In the Shadow of Man. (New York: Dell).

Greenewalt, C. A. (1967) Bird Song: Acoustics and Physiology. (Washington, D. C.: Smithsonian Institution).

Henke, W. L. (1966) Dynamic articulatory model of speech production using computer simulation. Unpublished Ph.D. dissertation, M.I.T., Cambridge, Mass., Appendix B.

Hewes, G. W. (1971) Language Origins: A Bibliography. (Boulder, Colo.: University of Colorado, Dept. of Anthropology).

Hollien, H., P. Moore, R. W. Wendahl, and J. F. Michel. (1966) On the nature of vocal fry. J. Speech Hearing Res. 9, 245-247.
Howells, W. W. (1968) Mount Carmel man: morphological relationships. In Proceedings, VIIIth International Congress of Anthropological and Ethnological Sciences, Vol. I, Anthropology (Tokyo).
Irwin, O. C. (1948) Infant speech: development of vowel sounds. J. Speech Hearing Dis. 13, 31-34.
Jakobson, R., C. G. M. Fant, and M. Halle. (1952) Preliminaries to Speech Analysis. (Cambridge, Mass.: MIT Press).
Kelemen, G. (1948) The anatomical basis of phonation in the chimpanzee. J. Morphology 82, 229-256.
Kempelen, W. R. von. (1791) Mechanismus der menschichen Sprache nebst der Beschreibung seiner sprechenden Maschine. (Vienna: J. B. Degen).
Kenyon, K. W. (1969) The Sea Otter in the Eastern Pacific Ocean. (Washington, D. C.: U. S. Government Printing Office).
Kimura, D. (1964) Left-right differences: the perception of melodies. Quart. J. Exp. Psychol. 16, 355-358.
Kirchner, J. A. (1970) Pressman and Kelemen's Physiology of the Larynx, rev. ed. (Rochester, Minn.: American Academy of Ophthalmology and Otolaryngology).
Kuipers, A. H. (1960) Phoneme and Morpheme in Kabardian. (The Hague: Mouton).
Ladefoged, P. and D. E. Broadbent. (1957) Information conveyed by vowels. J. Acoust. Soc. Amer. 29, 98-104.
Lane, H. (1965) Motor theory of speech perception: a critical review. Psychol. Rev. 72, 275-309.
Lenneburg, E. H. (1967) Biological Foundations of Language. (New York: Wiley).
Liberman, A. M., F. S. Cooper, D. P. Shankweiler, and M. Studdert-Kennedy. (1967) Perception of the speech code. Psychol. Rev. 74, 431-461.
Lieberman, P. (1963) Some acoustic measures of the periodicity of normal and pathologic larynges. J. Acoust. Soc. Amer. 35, 344-353.
Lieberman, P. (1967) Intonation, Perception, and Language. (Cambridge, Mass.: MIT Press).
Lieberman, P. (1968) Primate vocalizations and human linguistic ability. J. Acoust. Soc. Amer. 44, 1574-1584.
Lieberman, P. (1970) Towards a unified phonetic theory. Ling. Inq. 1, 307-322.
Lieberman, P. (1972) The Speech of Primates. (The Hague: Mouton).
Lieberman, P. and E. S. Crelin. (1971) On the speech of Neanderthal man. Ling. Inq. 2, 203-222.
Lieberman, P., E. S. Crelin, and D. H. Klatt. (1972a) Phonetic ability and related anatomy of the newborn, adult human, Neanderthal man and the chimpanzee. Amer. Anthropologist 74, 287-307.
Lieberman, P., K. S. Harris, P. Wolff, and L. H. Russell. (1972b) Newborn infant cry and nonhuman primate vocalizations. J. Speech Hearing Res. 14, 718-727.
Lieberman, P., D. H. Klatt, and W. A. Wilson. (1969) Vocal tract limitations on the vowel repertoires of rhesus monkey and other nonhuman primates. Science 164, 1185-1187.
Lindblom, B. and J. Sundberg. (1969) A quantitative model of vowel production and the distinctive features of Swedish vowels. Speech Transmission Laboratory Report 1 (Stockholm, Sweden: Royal Institute of Technology).
Lisker, L. and A. S. Abramson. (1964) A cross-language study of voicing in initial stops: acoustical measurements. Word 20, 384-422.

Lynip, A. W. (1951) The uses of magnetic devices in the collection and analysis of the preverbal utterances of an infant. Genetic Psychol. Monog. 44, 221-262.

Manly, R. S. and L. C. Braley. (1950) Masticatory performance and efficiency. J. Dent. Res. 29, 448-462.

Manly, R. S. and F. R. Shiere. (1950) The effect of dental deficiency on mastication and food preference. Oral Surg., Oral Med., Oral Path. 3, 674-685.

Manly, R. S. and P. Vinton. (1951) A survey of the chewing ability of denture wearers. J. Dent. Res. 30, 314-321.

Miller, J. M., D. Sutton, B. Pfingst, A. Ryan, and R. Beaton. (1972) Single cell activity in the auditory cortex of rhesus monkeys: behavioral dependency. Science 177, 449-451.

Montagu, M. F. A. (1962) Time morphology and neoteny in the evolution of man. In Culture and the Evolution of Man, ed. by M. F. A. Montagu. (New York: Oxford University Press) 324-342.

Morse, P. (1971) Speech perception in six-week old infants. Unpublished Ph.D. dissertation, University of Connecticut.

Negus, V. E. (1949) The Comparative Anatomy and Physiology of the Larynx. (New York: Hafner).

Nett, E. G. (in press) A note on phonetic ability. Amer. Anthropologist.

Nottebohm, F. (1970) Ontogeny of bird song. Science 167, 950-956.

Patte, E. (1955) Les Néanderthaliens, Anatomie, Physiologie, Comparaisons. (Paris: Masson).

Perkell, J. S. (1969) Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study. (Cambridge, Mass.: MIT Press).

Peterson, G. E. and H. L. Barney. (1952) Control methods used in a study of the vowels. J. Acoust. Soc. Amer. 24, 175-184.

Premack, D. (1972) Language in chimpanzee? Science 172, 808-822.

Rand, T. C. (1971) Vocal tract size normalization in the perception of stop consonants. Haskins Laboratories Status Report on Speech Research SR-25/26, 141-146.

Schultz, A. H. (1944) Age changes and variability in gibbons. Amer. J. Phys. Anthropol. n.s. 2, 1-129.

Schultz, A. H. (1955) The position of the occipital condyles and of the facet relative to the skull base in primates. Amer. J. Phys. Anthropol. n.s. 13, 97-120.

Schultz, A. H. (1968) The recent hominoid primates. In Perspectives of Human Evolution I, ed. by S. L. Washburn and P. C. Jay. (New York: Holt, Rinehart and Winston) 122-195.

Shankweiler, D. and M. Studdert-Kennedy. (1967) Identification of consonants and vowels presented to left and right ears. Quart. J. Exp. Psychol. 19, 59-63.

Stevens, K. N. (in press) Quantal nature of speech. In Human Communication, A Unified View, ed. by E. E. David and P. B. Denes. (New York: McGraw Hill).

Stevens, K. N. and A. S. House. (1955) Development of a quantitative description of vowel articulation. J. Acoust. Soc. Amer. 27, 484-493.

Straus, W. L., Jr. and A. J. E. Cave. (1957) Pathology and posture of Neanderthal man. Quart. Rev. Biol. 32, 348-363.

Troubetzkoy, N. S. (1939) Principes de Phonologie. (Paris, Klincksieck, 1949. Trans. by J. Cantineau).

Truby, H. M., J. F. Bosma, and J. Lind. (1965) Newborn Infant Cry. (Uppsala: Almqvist and Wiksells).

Van den Berg, J. W. (1960) Vocal ligaments versus registers. Current Problems in Phoniatrics and Logopedics 1, 19-34.

Vlček, E. (1970) Étude comparative onto-phylogénétique de l'enfant du Pech-de-l'Azé par rapport a d'autres enfants néanderthaliens. In L'enfant du Pech-de-Azé, ed. by D. Feremback et al. (Paris: Masson) 149-186.

Wollberg, Z. and J. D. Newman. (1972) Auditory cortex of squirrel monkey: response patterns of single cells to species-specific vocalizations. Science 175, 212-214.

Zhinkin, N. I. (1963) An application of the theory of algorithms to the study of animal speech--methods of vocal intercommunication between monkeys. In Acoustic Behavior of Animals, ed. by R. G. Busnel. (Amsterdam: Elsevier).

II. PUBLICATIONS AND REPORTS

III. APPENDIX

PUBLICATIONS AND REPORTS

Publications and Manuscripts

Research on Audible Outputs of Reading Machines for the Blind. F. S. Cooper, J. H. Gaitenby, I. G. Mattingly, P. W. Nye, and G. N. Sholes. Bulletin of Prosthetics Research (Spring 1972) BPR 10-17, 252-257.

Machines and Speech. Franklin S. Cooper. In Research Trends in Computational Linguistics, Report of a Conference. (Arlington, Va.: Center for Applied Linguistics, 1972).

Speech of Primates. Philip Lieberman. (The Hague: Mouton, 1972).

The following three papers appeared in Papers in Linguistics and Phonetics to the Memory of Pierre Delattre, ed. by Albert Valdman. (The Hague: Mouton, 1972):

Tonal Experiments with Whispered Thai.
Arthur S. Abramson, 21-30.

In Search of the Acoustic Cues.
Alvin M. Liberman and Franklin S. Cooper, 329-338.

Stop Duration and Voicing in English.
Leigh Lisker, 339-344.

Voice-Timing Perception in Spanish Word-Initial Stops. A. S. Abramson and Leigh Lisker, Journal of Phonetics (1973) 1, 1-8.

On Peripheral and Central Processes in Vision: Inferences from an Information-Processing Analysis of Masking with Patterned Stimuli. M. T. Turvey. Psychological Review (January 1973) 80, 1-52.

Auditory and Phonetic Memory Codes in the Discrimination of Consonants and Vowels. David Pisoni. Perception and Psychophysics (April 1973) 13, 253-260.

Laryngeal Control in Vocal Attack: An Electromyographic Study. H. Hirose and T. Gay. Folia Phoniatrica, in press.

The following two papers by Philip Lieberman will be published in the proceedings of the IX International Congress of Anthropological and Ethnological Sciences. (Chicago, Ill., September 1973):

* On the Evolution of Language: A Unified View.

* Linguistic and Paralinguistic Interchange.

---

*Appears in this report, SR-33.

Reply to "A Note on Phonetic Ability." P. Lieberman, E. S. Crelin, and D. H. Klatt. American Anthropologist, in press.

Visual Storage or Visual Masking?: An Analysis of the "Retroactive Contour Enhancement" Effect. M. T. Turvey, Claire Farley Michaels, and Diane Kewley Port. Quarterly Journal of Experimental Psychology, in press. (Also appeared in SR-31/32, 1972.)

*Are You Asking Me, Telling Me, or Talking to Yourself? Kerstin Hadding and Michael Studdert-Kennedy.

*Computer Processing of EMG Signals at Haskins Laboratories. Diane Kewley-Port.

*Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results). P. W. Nye and J. H. Gaitenby.

*Dichotic Release from Masking for Speech. Timothy C. Rand.

*Discrimination of Intensity Differences on Formant Transitions In and Out of Syllable Context. M. F. Dorman.

*Effect of Speaking Rate on Labial Consonant-Vowel Articulation. T. Gay, T. Ushijima, H. Hirose, and F. S. Cooper.

*Effects of Proactive Interference and Rehearsal on the Primary and Secondary Components of Short-Term Retention. M. T. Turvey and Robert A. Weeks

*On the Short-Term Retention of Serial, Tactile Stimuli. Edie V. Sullivan and M. T. Turvey.

*Perception of Speech and Nonspeech, with and without Transitions. James E. Cutting.

*Pitch Determination by Adaptive Autocorrelation Method. Georgije Lukatela.

*Phonetic Activity in Reading: An Experiment with Kanji. Donna Erickson, Ignatius G. Mattingly, and Michael T. Turvey.

*Phonological Fusion in Synthetic and Natural Speech. James E. Cutting.

*Speech Misperception: Inferences About a Cue for Cluster Perception from a Phonological Fusion Task. James E. Cutting.

*A Speech Perception Paradox?: The Right-Ear Advantage and the Lag Effect. Robert A. Weeks.

Reports and Oral Presentations

Knowing About Things You Don't Know You Know About. M. T. Turvey. Medical Research Center, Naval Submarine Base, New London, Conn., September 1972.

Constructivism, Perceptual Systems, and Tacit Knowledge. M. T. Turvey. Conference on Cognition and the Symbolic Processes. Pennsylvania State University, October 1972.

Activity of Some Extrinsic and Intrinsic Tongue Muscles in the Articulation of American-English Vowels. Larry Raphael and Katherine S. Harris. Acoustical Society of America, Miami Beach, Fla., November 1972.

*The Role of the Extrinsic and Intrinsic Tongue Muscles in Differentiating the English Tense-Lax Vowel Pairs. Lawrence J. Raphael and Fredericka Bell-Berti. Presented at the American Speech and Hearing Convention, San Francisco, Calif., November 1972.

A Lecture on Lecturing. M. T. Turvey. Workshop on College Teaching, University of Connecticut, Storrs, November 1972.

On the Evolution of Human Language. Philip Lieberman. Queens College, City University of New York; and Columbia University, N. Y., December 1972.

The Phonetic Reality of Categorical Labels. Larry Raphael. Speech Communication Association, Chicago, Ill., December 1972.

*An Electromyographic Study of the American English Liquids. David R. Leidner. Presented at the annual meeting of the Linguistic Society of America, Atlanta, Ga., December 1972.

*Cross-Language Study of the Perception of the F3 Cue for [r] versus [l] in Speech- and Nonspeech-Like Patterns. Kuniko Miyawaki, A. M. Liberman, O. Fujimura, Winifred Strange, and J. J. Jenkins. Presented at the International Congress of Psychology, Tokyo, 1972.

Studies in Short-Term Memory; and An Information-Processing Analysis of Masking. M. T. Turvey. Two Talks given at Rockefeller University, N. Y., January 1973.

Levels of Processing in Language Perception and Production. Ruth S. Day. Invited address, New York Academy of Sciences, 7 February 1973.

Language Development: Loss of Pattern Discrimination? Ruth S. Day. International Neurophysiological Society, Symposium on "Developmental Cerebral Dominance." New Orleans, La., 9 February 1973.

Cracking the Phonetic Code. M. Studdert-Kennedy. Psycholinguistic Circle of New York, 13 February 1973.

Implications of Cognitive Psychology for the Teaching-Learning Process. Ruth S. Day. Invited address, Title IV Desegregation Project, Center for Urban Education, University of Nebraska, Omaha, 17 March 1973.

Colloquia. Ruth S. Day. University of Toronto, 31 January 1973; The Rockefeller University, 1 March 1973; University of Nebraska, Omaha, 16 March 1973.

*Segmentation of the Spoken Word and Reading Acquisition. Isabelle Y. Liberman. Presented at the meeting of the Society for Research in Child Development, Philadelphia, Pa., 31 March 1973.

Knowing More Than We Know. M. T. Turvey. Engineering Honors Societies, University of Connecticut, Storrs, April 1973.

Two Central Operations in Perceiving at a Single Glance. M. T. Turvey. Brown University, Providence, R. I., April 1973.

*Forward and Backward Masking of Brief Vowels. M. Dorman, D. Kewley-Port, S. Brady-Wood, and M. T. Turvey. Presented at the 85th meeting of the Acoustical Society of America, Boston, Mass., April 1973.

Digit Span Memory in Language-Bound and Stimulus-Bound Subjects. Ruth S. Day. Presented at the 85th meeting of the Acoustical Society of America, Boston, Mass., 11 April 1973.

A Two-Pass Procedure for Synthesis by Rule. Gary Kuhn. Presented at the meeting of the Acoustical Society of America, Boston, Mass., April 1973.

APPENDIX

DDC (Defense Documentation Center) and ERIC (Educational Resources Information Center) numbers:

SR-21/22 to SR-31/32

Status Report		DDC	ERIC
SR-21/22	January - June 1970	AD 719382	ED-044-679
SR-23	July - September 1970	AD 723586	ED-052-654
SR-24	October - December 1970	AD 727616	ED-052-653
SR-25/26	January - June 1971	AD 730013	ED-056-560
SR-27	July - September 1971	AD 749339	ED-071-533
SR-28	October - December 1971	AD 742140	ED-061-837
SR-29/30	January - June 1972	AD 750001	ED-071-484
SR-31/32	July - December 1972	AD 757954	

AD numbers may be ordered from: U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, Virginia 22151

ED numbers may be ordered from: ERIC Document Reproduction Service
Leasco Information Products, Inc.
P. O. Drawer O
Bethesda, Maryland 20014

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
Haskins Laboratories, Inc. 270 Crown Street New Haven, Connecticut 06510	Unclassified 2b. GROUP N/A

3. REPORT TITLE

Status Report on Speech Research, no. 33, January-March 1973

4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)

Interim Scientific Report

5. AUTHOR(S) (First name, middle initial, last name)

Staff of Haskins Laboratories; Franklin S. Cooper, P.I.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
May 1973	286	385

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
ONR Contract N00014-67-A-0129-0001-0002 NIDR: Grant DE-01774 NICHD: Grant HD-01994 NIH/DRFR: Grant RR-5596 NSF: Grant GS-28354 VA/PSAS Contract V101(134)P-71 NICHD Contract NIH-71-2420 The Seeing Eye, Inc. Equipment Grant	SR-33 (1973) 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) None

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited.*

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
N/A	See No. 8

13. ABSTRACT This report (1 January-31 March) is one of a regular series on the status and progress of studies on the nature of speech, instrumentation for its investigation, and practical applications. Manuscripts cover the following topics:

-Linguistic and Psychophysical Judgments of Fundamental Frequency Contours
-Discrimination of Intensity Differences on Formant Transitions
-Phonological Fusion in Synthetic and Natural Speech
-Right-Ear Advantage and the Lag Effect
-Perception of Speech and Nonspeech, with and without Transitions
-Dichotic Release from Masking for Speech
-Speech Misperception: Phonological Fusion Task
-Cross-Language Study of Perceptual Cues for [r] and [l]
-Consonant Intelligibility in Synthetic and Natural Speech
-Forward and Backward Masking of Brief Vowels
-Short-Term Retention: Effects of Proactive Interference and Rehearsal
-Short-Term Retention of Serial, Tactile Stimuli
-Phonetic Activity in Reading: Experiment with Kanji
-Segmentation of the Spoken Word and Reading Acquisition
-Linguistic and Paralinguistic Interchange
-Computer Processing of EMG Signals
-Pitch Determination by Adaptive Autocorrelation Method
-Electromyographic Study of American English Liquids
-Differentiation of Tense-Lax Vowels: EMG Study of Tongue Muscles
-Effect of Speaking Rate on Labial Consonant-Vowel Articulation
-Evolution of Language

DD FORM 1 NOV 65 1473 (PAGE 1)

S/N 0101-807-6811

*This document contains no information not freely available to the general public. It is distributed primarily for library use.

UNCLASSIFIED

Security Classification

A-31408

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Auditory perception						
Speech perception						
Dichotic fusion						
Dichotic speech perception						
Self-masking of speech						
Synthetic speech intelligibility						
Speech masking						
Short-term memory						
Short-term memory for touch						
Reading: phonetic component						
Reading: kanji						
Reading: phonetic activity						
Linguistics and paralinguistics						
EMG: speech						
EMG: speaking rate						
EMG: computer processing						
Pitch by autocorrelation						
Evolution of language						